International Journal on Advances in Systems and Measurements

2009 vol. 2 nr. 4

The *International Journal on Advances in Systems and Measurements* is published by IARIA.
ISSN: 1942-261x
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Systems and Measurements
Volume 2, Number 4, 2009

- Igor Podebrad, Commerzbank, Germany
- Leon Reznik, Rochester Institute of Technology, USA
- Chi Zhang, Juniper Networks, USA

Sensor Networks

- Steven Corroy, University of Aachen, Germany
- Mario Freire, University of Beira Interior, Portugal / IEEE Computer Society - Portugal Chapter
- Jianlin Guo, Mitsubishi Electric Research Laboratories America, USA
- Zhen Liu, Nokia Research – Palo Alto, USA
- Winston KG Seah, Institute for Infocomm Research (Member of A*STAR), Singapore
- Radosveta Sokkulu, Ege University - Izmir, Turkey
- Athanasios Vasilakos, University of Western Macedonia, Greece

Electronics

- Kenneth Blair Kent, University of New Brunswick, Canada
- Josu Etxaniz Maranon, Euskal Herriko Unibertsitatea/Universidad del Pais Vasco, Spain
- Mark Brian Josephs, London South Bank University, UK
- Michael Hubner, Universitaet Karlsruhe (TH), Germany
- Nor K. Noordin, Universiti Putra Malaysia, Malaysia
- Arnaldo Oliveira, Universidade de Aveiro, Portugal
- Candid Reig, University of Valencia, Spain
- Sofiene Tahar, Concordia University, Canada
- Felix Toran, European Space Agency/Centre Spatial de Toulouse, France
- Yousaf Zafar, Gwangju Institute of Science and Technology (GIST), Republic of Korea
- David Zammit-Mangion, University of Malta-Msida, Malta

Testing and Validation

- Cecilia Metra, DEIS-ARCES-University of Bologna, Italy
- Krzysztof Rogoz, Motorola, Poland
- Rajarajan Senguttuvan, Texas Instruments, USA
- Sergio Soares, Federal University of Pernambuco, Brazil
- Alin Stefanescu, SAP Research, Germany
- Massimo Tivoli, Universita degli Studi dell'Aquila, Italy

Simulations

- Tejas R. Gandhi, Virtua Health-Marlton, USA
- Ken Hawick, Massey University - Albany, New Zealand
- Robert de Souza, The Logistics Institute - Asia Pacific, Singapore
- Michael J. North, Argonne National Laboratory, USA

Additional reviews by:

- Tiago Massoni, UFCG, Brazil

International Journal on Advances in Systems and Measurements
Volume 2, Number 4, 2009

CONTENTS

System Level Analysis for Achieving Thermal Balance and Lifetime Reliability in Reliably Overclocked Systems **258 - 268**
Prem Kumar Ramesh, Iowa State University, USA
Viswanathan Subramanian, Iowa State University, USA
Arun K. Somani, Iowa State University, USA

Biodiversity Information Systems Evolution: The MABIS model to gather several communities on an adaptable environment **269 - 282**
Didier Sébastien, Université de la Réunion, France
Noël Conruyt, Université de la Réunion, France
Rémy Courdier, Université de la Réunion, France
Nicolas Sébastien, Université de la Réunion, France
Tullio Tanzi, Institut TELECOM - TELECOM ParisTech, France

ASSOLO: an Efficient Tool for Active End-to-end Available Bandwidth Estimation **283 - 292**
Emanuele Goldoni, University of Pavia, Italy
Giuseppe Rossi, University of Pavia, Italy
Alberto Torelli, University of Pavia, Italy

System Level Analysis for Achieving Thermal Balance and Lifetime Reliability in Reliably Overclocked Systems

Prem Kumar Ramesh, Viswanathan Subramanian and Arun K. Somani
Dependable Computing and Networking Laboratory
Iowa State University
Ames, IA, USA
{pramesh, visu, arun}@iastate.edu

***Abstract*—Advancements in process technology offer continuous improvements in system performance. Technology scaling brings forth several new challenges. In particular, process, voltage, and temperature variations require sufficient safety margins to be added to the clock frequency of digital systems, making it overly conservative. Aggressive, but reliable, dynamic clock frequency tuning mechanisms that achieve higher system performance, by adapting the clock rates beyond worst-case limits, have been proposed earlier. Even though reliable overclocking guarantees functional correctness, it leads to higher power consumption and overheating. As a consequence, reliable overclocking without considering on-chip temperatures will bring down the lifetime reliability of the chip. In [1], we presented a comparative study on the thermal behavior of reliably overclocked systems with non-accelerated systems. In this paper, we elaborate more on the theoretical analysis along with experimental results to establish a safe acceleration zone for such 'better than worst-case' designs by efficiently balancing the gains of overclocking and the impact on system temperature. We analyze how reliable overclocking impacts the on-chip temperature of microprocessors, and evaluate the effects of overheating, due to reliable dynamic overclocking mechanisms, on the lifetime reliability of such systems. First, we theoretically study the possibilities for realizing such a system. We, then, evaluate the effects of thermal throttling, a technique that clamps the on-chip temperature below a predefined value, on system performance and reliability. Our study shows that a reliably overclocked system with dynamic thermal throttling, constrained to operating within 355K, achieves around 25% performance improvement.**

***Index Terms*—Microprocessors, Dependability, Adaptability, Overclocking, Thermal Throttling**

I. Introduction

In pursuit of ever faster execution times, a growing community known as overclockers, are manually accelerating their high performance processors past the manufacturer specified limits. Impressive results have been shown in existing systems that support overclocking. For example, a 2.6 GHz AMD Phenom processor running at speeds of up to 4 GHz using liquid cooling has been achieved. Such is the interest with overclocking enthusiasts that chipset manufacturers are introducing technologies that support overclocking. AMD's Overdrive and Advance Clock Calibration technologies are cases in point [2]. These gains are possible because of the worst-case assumptions used by traditional design methodologies. The clock frequency of a processor is selected to give enough time for the longest delay path, which determines the circuit propagation delay, to stabilize under adverse operating conditions. This propagation delay varies as process, voltage and temperature (PVT) variations are introduced during circuit fabrication and operation; so designers must assume the worst when fixing the system clock frequency. However, the combination of longest delay paths and adverse operating conditions are rare, leading to room for performance improvement that overclockers exploit. Systems running at overclocked speeds cannot be relied upon, as the possibility of system failure is very high. As a result, to account for the timing errors that occur at better-than-worst-case speeds, it is important to overclock the system reliably, in order to reap the benefits of making the common case faster.

The design for worst-case settings provides us an opportunity to improve processor performance to a greater extent through overclocking. When the system is forced to operate beyond this conservative limit, reliable overclocking mechanisms employ proven fault tolerance techniques to detect and recover from timing errors. Although aggressive clocking mechanisms facilitate in improving performance, they adversely impact on-chip temperatures, leading to hot spots. Overclocking enthusiasts invest heavily in expensive cooling solutions to protect the chip from overheating, and such overclocked systems typically have significantly lower lifetime. Additionally, reliable overclocking techniques necessitate additional circuitry, leading to an increase in power consumption. Higher clock speeds and power densities invariably lead to accretion of on-chip temperature over a period of time. As systems operate faster, on-chip temperatures quickly reach and exceed the safe limits. This poses a serious threat to the lifetime reliability of these systems. In [1], we presented our comparative study on the thermal behavior of reliably overclocked systems with non-accelerated systems. In this paper, we elaborate more on the theoretical analysis along with experimental results to establish a safe acceleration zone for such 'better than worst-case' designs by efficiently balancing the gains of overclocking and the impact on system temperature.

We must emphasize that current products from both the leading microprocessor vendors, Intel and AMD, have dynamic thermal monitoring techniques that take necessary corrective action to maintain on-chip temperature [3]–[5]. The corrective actions, in most cases, shut down the system or reduce system voltage and frequency, leading to considerable performance degradation. Our goal in this study is to analyze the temperature pattern of reliably overclocked systems, and

evaluate the lifetime reliability of such reliable aggressive clocking mechanisms. Furthermore, we monitor the on-chip temperature of aggressively overclocked systems that dynamically enhance single threaded application performance. We couple thermal monitoring techniques with reliable overclocking to alleviate lateral issues relating to system power and reliability. While taking feedback from an integrated thermal monitor, we observed an average performance increase of 25%, while operating within temperature 355K. Our work is related to Razor [6], which uses timing error tolerance mechanism to conserve energy while suffering moderate performance degradation. Another relevant study, SPRIT^{3}E [7], uses a similar mechanism to reliably overclock the system to enhance system performance.

First, we theoretically analyze the possibilities for realizing a controlled reliably overclocked system, while maintaining the on-chip temperature. We use SimpleScalar [8] simulator for Alpha EV6 processor, with a built-in power model, namely, Wattch [9] for the experiments. We integrate HotSpot [10] thermal modeling tool to monitor on-chip temperature. In real hardware, this translates to thermal sensors and counters for tracking on-chip temperature, which most of the present day chips support. We explore a broad spectrum of results for SPEC 2000 benchmark suite [11].

The remainder of this paper is organized as follows. Section II provides an overview of how reliable overclocking is performed in processors for performance enhancement. This section also outlines the issues related to thermal and reliability management in processors. We use processors and systems interchangeably in the rest of the paper. Section III explains our experimental framework used for analyzing the thermal impacts in reliably overclocked processors. We present our results in Section IV and Section V concludes the paper.

II. Background

A. Reliable Overclocking

One of the earliest works on aggressive clocking, TEATIME [12], scales the frequency of a pipeline using dynamic timing error avoidance. This technique attempts to achieve better-than-worst-case performance by realizing typical delay operation rather than assuming worst-case delays and operating conditions. TEATIME achieves this by modeling a one-bit wide delay chain that reflects the worst-case critical path of the system, plus a safety margin. A prior work to this called TIMERTOL [13] exists in which, timing error tolerance is achieved by multiple special copies of the pipeline logic. Similar architectures include CTV [14] and X-Pipe [15] that propose timing speculation at pipeline stage level.

The most significant aspect that can be exploited by reliable overclocking is the input data dependency of the worst-case delays. The worst-case delay paths are sensitized only for specific input combinations and data sequences [16]. Typically, the propagation delay of the digital system is much less than the worst-case delay and this can be exploited by overclocking. The benefits of overclocking can be furthered by allowing a tolerable number of errors to occur, and have an efficient mechanism to detect and recover from those errors. In addition to this, systems have different design restrictions, such as power, energy or area constraints. Based on all this, there are numerous architectures that have been proposed over the years.

Timing speculation based architectures that replicate registers in circuit critical path have been proposed. The basic idea is to duplicate latching, using shadow latches that are clocked in such a way to guarantee correctness. When a timing error is detected, it is recovered the following cycle. This technique along with dynamic voltage scaling has been used to improve energy efficiency [6]. Along with adaptive clocking mechanisms, reliable overclocking improves performance drastically [7].

In [17], the trade-off between reliability and performance is studied, and overclocking is used to improve the performance of register files. The conjoined pipeline architecture, proposed in [18], organizes pipeline redundancy in such a way as to improve both performance and reliability. In [19], triple modular redundancy based timing speculative register cells that can handle both soft errors and timing errors have been proposed.

Other works in the domain seek to improve common case performance through functionally incorrect designs [20], [21]. The Selective Series Duplex architecture [22] consists of an integrity checking architecture for superscalar processors that can achieve fault tolerance capability of a duplex system at much less cost than the traditional duplication approach. DIVA [21] uses spatial redundancy by providing a separate, slower pipeline processor alongside the fast processor. The desire for better than worst case designs is much more serious in nanoscale technology. PVT variations within and across the die are causing a bottleneck while selecting the worst-case frequency. ReCycle [23] uses additional registers and clock buffers to apply cycle time stealing from the faster pipeline stages to the slower ones. Another technique, EVAL [24] has been proposed to maximize performance with low power overhead in the presence of timing induced errors.

Apart from these run-time schemes, there are static methods that are specifically developed for better than worst case architectures. The effect of parameter variations and its impact on timing errors has been studied in [25]. BlueShift [26] proposes a design methodology from ground up. The main idea is to identify and optimize the frequently used critical paths, called the 'overshooters', at the expense of the lesser frequent ones.

In this section, we briefly discuss an in-built error detection and recovery mechanism that tolerates timing errors occurring at frequencies past the worst-case limit. We describe the working of *local timing error detection and recovery* circuits that replicate pipeline registers to support reliable overclocking. These circuits were first proposed in [6] to implement energy efficient processors and later used in [7] to enhance performance of superscalar processors. The purpose of these circuits is to detect and correct any resultant timing errors that occur because of overclocking, and to guarantee computational correctness.

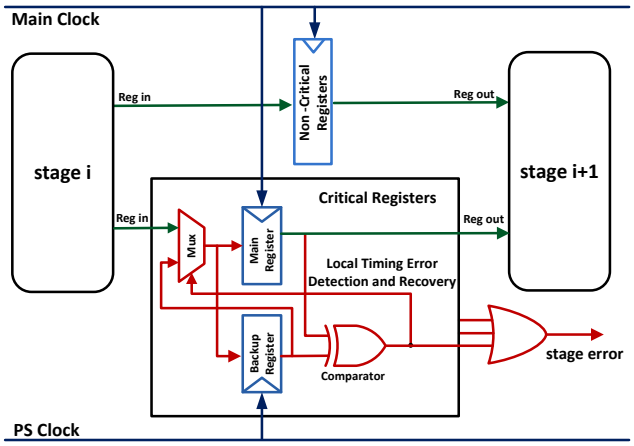

Fig. 1. Typical pipeline stage in a Reliably Overclocked Processor. Local timing error detection and recovery scheme for critical registers is shown in detail.

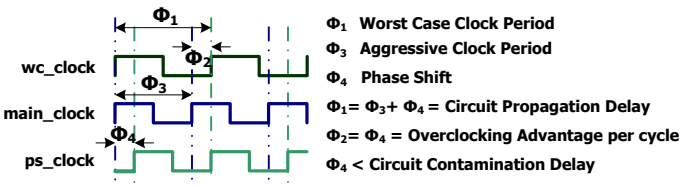

Fig. 2. Timing diagram showing overclocking advantage per cycle, as compared to the worst-case clock

B. Timing Error Detection and Recovery

In a reliably overclocked processor (ROP), to tolerate timing errors, registers in the critical paths of every pipeline stage are augmented with a second time-delayed register. A typical pipeline stage in such a processor, along with local timing error detection and recovery circuit augmentation for critical path registers, is shown in Figure 1. Each combinational logic stage is a dense logic combination with multiple inputs and outputs, and possibly with more than one path from each input to output. The short paths in the logic can operate correctly even during extreme voltage and/or frequency scaling. The paths that are not likely to meet their timing requirements are categorized as critical paths and only their corresponding stage output registers are replaced with timing error detection and recovery circuits.

A brief description of timing error detection and recovery in a ROP is presented from [7]: The main register is clocked ambitiously by the *Main Clock* at a frequency higher than that required for error-free operation. The backup register is clocked in such a way that it is prevented from being affected by timing errors, and its output is considered "golden" [7]. The clock for this register is phase shifted, shown as *PS Clock*, such that the combinational logic is effectively given its full, worst-case propagation delay time to execute. In case of a mismatch between the primary and backup registers, a recovery measure is taken by correcting the current stage data and stalling the pipeline for one cycle. In addition to local recovery, action is also taken on a global scale to maintain correct execution of the pipeline in the event of a timing error. The extent to which systems can be overclocked is limited by the penalty cycles needed to recover from timing errors. A balance must be maintained between the number of cycles lost to error recovery and the gains of overclocking. The achievable performance enhancement per cycle is shown in Figure 2 as Φ_2.

C. Timing Error Based Feedback Control System

Reliably overclocking a processor may not yield an increase in performance at all times. The reason being that the occurrence of a timing error is highly dependent on the workload and the current operating conditions. Therefore, it is beneficial to have an adaptive clock tuning system, which increases or decreases the clock frequency based on a set target error rate. In other words, it is necessary to fix a bound for overclocking, as errors induce additional recovery cycles that imparts a performance overhead.

- Let t_{no} denote the non-overclocked worst-case time period and t_{ov} denote the time period after overclocking.
- Let t_{diff} be the difference in time between the two time periods. Then, to execute n clock cycles, the total execution time is reduced by $t_{diff} \times n$, when there is no error.
- Let S_e, k and t_{pll} denote the fraction of clock cycles affected by errors, error recovery cycles and time taken by Phase Locked Loop (PLL) to lock next frequency respectively.

Then, equation 1 gives the bound on the timing errors that can be tolerated without adding any performance penalty.

$$S_e < \frac{t_{diff}}{t_{ov} \times k} - \frac{t_{pll}}{n \times t_{ov} \times k} \quad (1)$$

Dynamic clock frequency tuning is controlled by a global feedback system based on the total number of timing errors that occur in a specified time interval. Current products, such as IBM PowerPC 750GX processors, use two PLL scheme for clock generation to perform dynamic power-performance scaling [27]. This allows instant frequency switching, when frequency sampling interval is greater than t_{pll}.

The number of errors occurring at each timing error counter sampling interval is continuously monitored. As long as the number of errors is within target limits, the frequency is scaled up, else scaled down. One can apparently construe that the error rate is a monotonically increasing function with respect to frequency. This allows the use of efficient search algorithms to select the next tuned frequency starting from the base frequency.

For our understanding, let us consider the empirical model for circuit delay as given by Eqn (2).

$$Delay = \frac{C.V^2}{2v_{SAT}C_{OX}W(V - V_T)^2} \quad (2)$$

Here, v_{SAT}, C_{OX} and W are technology dependent constants; C specifies the capacitive load the circuit drives; V and V_T

are the system voltage and threshold voltage respectively ($V_T = 0.2398V$ for $45nm$ technology) [28]. Eqn (2) suggests that the time period provided should match this *Delay* in order to avoid timing errors. In traditional designs, the clock frequency is determined off-line during design phase for the worst-case settings, which is too conservative. However in our case, the clock time period is shorter compared to the circuit delay, and this results in timing errors. That is, there is a direct correspondence between the number of timing errors and the circuit slow down. Further, the slowdown can be related to the capacitive load that can be driven for that time period. Rearranging Eqn (2) for t_{no} and t_{ov} yield the following loads that can be driven respectively.

$$K_{no} = \frac{t_{no}(V-V_T)^2}{V^2},\ K_{ov} = \frac{t_{ov}(V-V_T)^2}{V^2}$$

Thus, the percentage slow down for the overclocked frequency with respect to the worst-case frequency is given by Eqn 3.

$$\%SlowDown = \frac{K_{no} - K_{ov}}{K_{no}} \times 100 \tag{3}$$

Finally, the maximum frequency for performance enhancement is theoretically limited by the contamination delay of the circuit. If time period of the new frequency is less than contamination delay of the circuit, timing error certainly occurs during every cycle and the error rate goes to 100%. Frequencies below propagation delay do not cause any timing errors at all (0% error rate). This, however, incurs performance overhead. Earlier studies have indicated that fixing a non-zero target error rate improves performance significantly. However, the system temperature goes up along with the system performance as the target error rate margin increases.

1) Speed-up: Having discussed the importance of timing error throttling, the next step is to re-calculate the speed-up achieved from overclocking. In a pipelined processor, the total number of cycles to process a given set of instructions is mainly divided into instruction execution, branch penalty and memory cycles. During overclocking, the clock frequency of the memory is not scaled, thereby increasing the total number of execution cycles. Let each memory operation take C_m cycles at t_{no} and q be the factor by which the frequency is scaled i.e., ($q = \frac{t_{no}}{t_{ov}}$). Now, after overclocking each memory operation takes $q.C_m$ cycles.

Let us assume that the system takes n clock cycles without considering memory cycles. If α denotes the factor of memory accesses that happen when the system executes n cycles. Then, the new execution time due to reliable overclocking is given by:

$$Ex_{ov} = n.t_{ov} + n.\alpha.q.C_m.t_{ov} + n.S_e.k.t_{ov} \tag{4}$$

To express original runtime (Ex_{no}) from Eqn 4, we replace t_{ov} by t_{no} and substitute $q = 1$ & $S_e = 0$. The overall speed up is calculated as given by Eqn 5.

$$Speedup = \frac{Ex_{no}}{Ex_{ov}} = \frac{q \times (1 + \alpha.1.C_m)}{(1 + \alpha.q.C_m + S_e.k)} \tag{5}$$

In our case, we take one cycle for timing error recovery, that is $k = 1$.

It should be noted that the benefits of reliable overclocking surpass the memory penalties provided the error rate is limited. This is clearly understood from the series of charts (a), (b) and (c) of Figure 3. Here, we depict the speed-up for a spectrum of memory access factors relative to target error rate, S_e, for different values of q.

For performance enhancement, the system must tolerate $20\%, 50\%, 70\%$ and 100% of timing exceptions at the overclocking rates $q = 1.2, 1.5, 1.7$ and 2.0, respectively. In the forthcoming sections, we show that for practical workloads the number of timing errors produced is quite low for smaller values of q, but quickly reaches 100% for higher values.

Also, from the model we deduce that non-memory bound workloads are more beneficial. For typical workloads, α is quite small, as most of the memory operations are shadowed through caching and buffering.

D. Thermal and Reliability Management

Over the last decade, thermal awareness has gained importance distinguishing itself from power awareness. Processor chips began to have thermal sensors in various locations to regularly sample the temperature and to shut down the operation in case of overheating. However, rapid heating and cooling of processor chips create thermal cycles affecting the lifetime reliability of the system [29].

The power consumed by a VLSI chip consists of two parts: dynamic and static. Dynamic power is dependent on capacitance (C), voltage (V), frequency (f) and switching factor (α), and is given by $P_{dyn} = \alpha CV^2 f$. Since dynamic power is directly proportional to the frequency at which the circuit operates, this causes overclocked systems to consume more power, which in turn causes systems to overheat. But solving the thermal problem is not as simple as bringing down the overall power consumed [30].

The problem becomes much more noticeable in designs under $90nm$ technology, where leakage power grows significantly. The leakage power grows exponentially with temperature as given by the empirical relationship, $P_{leak} \propto e^{\beta(T_i - T_0)}$ [31]. Here, β is technology dependent constant (β is 0.036 and 0.017 for $180nm$ and $70nm$ respectively), T_0 is the temperature of a reference point and T_i is the temperature at i^{th} instant with respect to the reference point. We see a positive feedback, wherein, increase in temperature leads to further leakage and increased total power consumption, which in turn leads to increase in temperature. Due to non-uniform switching and leakage, temperature is not distributed uniformly across the chip, creating localized heating in parts leading to hot spots.

Furthermore, overclocking increases the switching activity of the circuits causing more dynamic power dissipation. The dynamic power and energy consumed by the overclocked system are illustrated in Eqn (6) (7), respectively.

$$P_{ov} = \alpha.C.V_{ov}^2/t_{ov} = \alpha.C.V_{ov}^2/(t_{no}.q) \tag{6}$$

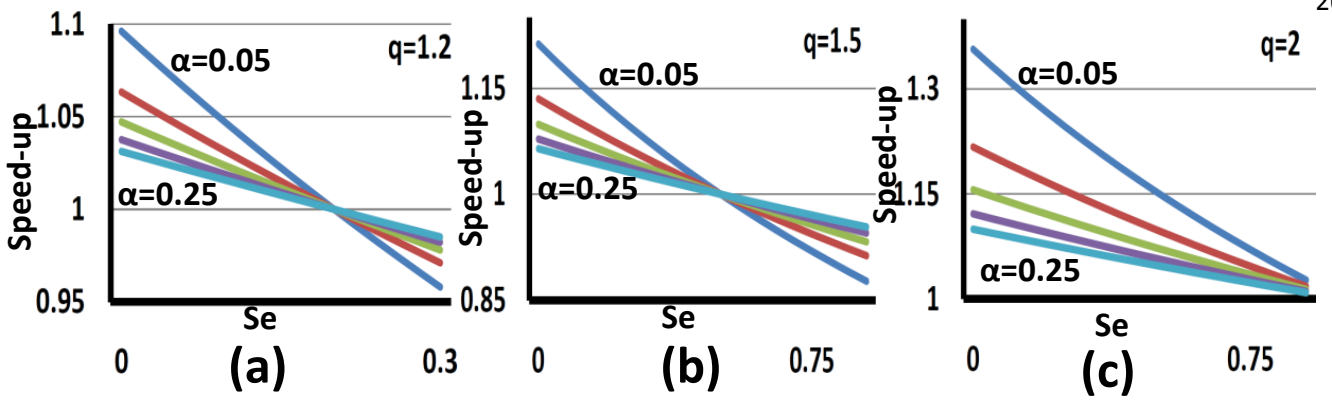

Fig. 3. Performance analysis of overclocking with error rate for different memory bounded workloads (a) Overclocking at 1.2X (b) Overclocking at 1.5X (c) Overclocking at 2X

$$E_{ov} = P_{ov}.n.(1 + \alpha.q.C_m + S_e.k).t_{ov} \quad (7)$$

Higher temperatures not only increase power budget, but also affect the lifetime reliability of the devices. To improve the overall reliability and lifetime of the systems, the thermal performance should be monitored and the average degradation of transistors managed. An initial exploration on thermal throttling through voltage reductions has been proposed in [32]. In this paper, we implement five critical failure mechanisms that are specified in [33] and [29] for our evaluation. A brief description of each of the wear out phenomenon and their respective Mean-Time-To-Failure (MTTF) are described below.

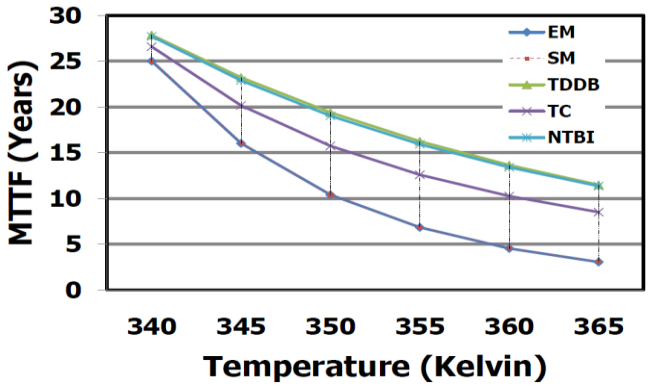

Fig. 4. MTTF for different steady state temperatures

Electromigration (EM) occurs due to transport of material due to gradual movement of the ions in a conductor caused by the momentum transfer between electrons and the diffusing metal. In Eqn (8), J is the interconnect current density. Activation energy, E_{aEM} and n are constants that depend on the interconnect metal used. (Typically, $n = 1.1$, $E_{aEM} = 0.9eV$).

$$MTTF_{EM} \propto (J)^{-n} e^{\frac{E_{aEM}}{kT}} \quad (8)$$

Stress Migration (SM) is a phenomenon that creates voids in the circuit, as a result of hydrostatic stress gradient. These voids may lead to high impedance or even break the circuit. This occurs due to difference in thermal expansion rates of materials. Again, E_{aSM}, m and the metal deposition temperature, T_{metal} are metal dependent constants in Eqn(9). T_{metal} generally assumes a value far higher than circuit operating temperature. (Typically, $m = 2.5$, $E_{aSM} = 0.9$). Although the term $|T_{metal} - T|^{-m}$ increases with T, there is an overall negative impact on the MTTF due to the exponential dependence of temperature.

$$MTTF_{SM} \propto |T_{metal} - T|^{-m} e^{\frac{E_{aSM}}{kT}} \quad (9)$$

Time Dependent Dielectric Breakdown (TDDB), also known as oxide breakdown occurs as a result of destruction of the gate oxide layer, and gradually leads to permanent transistor failure. a, b, X, Y and Z in Eqn (10) are fitting parameters. (Typically,$a = 78$, $b = -0.081$, $X = 0.759eV$, $Y = -66.8eV/K$, $Z = -8.37e - 4eV/K$).

$$MTTF_{TDDB} \propto (\frac{1}{V})^{(a-bT)} e^{\frac{[X+(Y/T)+ZT]}{kT}} \quad (10)$$

Sudden raise or fall in temperature causes **Thermal Cycles (TC)** which ultimately lead to device failure. Thermal cycles are caused by differences in thermal expansion rates across metal layers. Thermal cycling is proportional to the difference between current temperature and the ambient temperature $T_{ambient}$. In Eqn(11), $q = 2.35$, refers to the Coffin-Mason exponent, which is empirically determined material dependent constant. From this definition, one could observe that sudden cooling of devices below $T_{ambient}$ worsens the lifetime reliability.

$$MTTF_{TC} \propto (\frac{1}{T - T_{ambient}})^{q} \quad (11)$$

Finally, **Negative Bias Temperature Instability (NBTI)** is the failure mechanism that takes place in PFET devices.

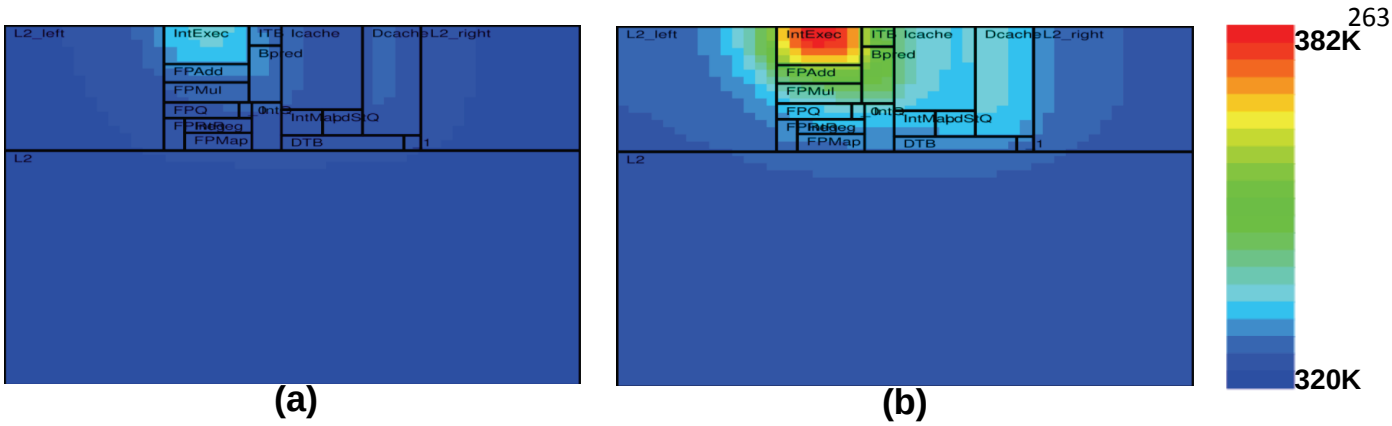

Fig. 5. Steady state temperature analysis (a) Non-overclocked processor settles around 330K (b) Reliably overclocked processor reaches over 380K

NBTI occurs due to timing constraint violations. In Eqn (12), A, B, C, D and β_1 are fitting parameters. (Typically, $A = 1.6328$, $B = 0.07377$, $C = 0.01$, $D = 0.06852$, $\beta_1 = 0.3$).

$$MTTF_{NBTI} \propto \left[\left\{ \begin{array}{c} ln\left(\frac{A}{1+2e^{B/kT}}\right) \\ -ln\left(\frac{A}{1+2e^{B/kT}} - C\right) \end{array} \right\} \times \frac{T}{e^{-D/kT}} \right]^{\beta_1} \quad (12)$$

Here, k is Boltzmann's constant and T is temperature in Kelvin. These wear out phenomena create impedance in the circuits, gradually leading to permanent device failures. Figure 4 shows how the increase in steady state temperature affects the processor lifetime. We use this reliability model to determine the critical temperature, for a target lifetime.

E. Thermal Consequences of Overclocking

Reliable dynamic clock frequency tuning for performance enhancement, described above, is incomplete without considering the thermal effects. Processors cannot be overclocked indefinitely, as this intensifies on-chip temperature. Thermal plots shown in Figure 5 compares a non-overclocked Alpha EV6 processor, running at $1GHz$ with an overclocked one, running at $2GHz$. We observed that steady state for dynamic reliable overclocking reached $380K$, while the non-overclocked settles at around $330K$. This calls the need for an efficient scheme for thermal balance in reliably overclocked processors.

III. Experimental Framework for Estimating On-chip Temperature

Figure 6 presents the entire simulation framework. The individual components are explained below in detail. The figure depicts both timing error based feedback control, and thermal throttle. For our initial evaluation of how on-chip temperatures vary when reliably overclocked, we only observe the temperature, without employing any thermal throttle. We employ dynamic clock tuning beyond worst-case limits, using timing error based feedback control, to adapt system behavior based on workload characteristics. The number of timing errors occurring at a given time is based on the workload being executed by the processor.

A. Modeling the ROP

To evaluate the trends in on-chip temperature, we model a reliably overclocked processor using a functional simulator, which incorporates a random timing error injector based on error profiles obtained by running application binaries on a hardware model. Our base processor, which is an out-of-order 64-bit, 4-way issue Alpha EV6 processor, is derived from the SimpleScalar-Alpha tool set [8]. This processor executes the Alpha AXP ISA. For our workload, we use pre-compiled set of Alpha binaries from the SPEC 2000 benchmark suite [11].

Wattch [9] is an accurate, architecture level power tool that is embedded within the SimpleScalar simulator. Wattch calculates instantaneous power at every cycle, and outputs the total power accumulated over a simulated period of time and the average power. We modified Wattch to track instantaneous power for each functional block.

B. Thermal Modeling

Thermal sensor modeling is done using the HotSpot tool [10]. The instantaneous power trace provided by Wattch power tool is used to calculate temperature. Since Wattch does not account for leakage power due to thermal runaway phenomenon, the temperature from HotSpot is used to calculate leakage power based on the formula presented in Section II. We obtain the $45nm$ Alpha EV6 floor plan from $130nm$ floor plan by assuming scaling is proportional to square of technology [34]. We also estimate the power dissipated by the additional circuitry required to detect timing errors.

C. Incorporating Timing Errors

In order to bring in the aspects of timing error in the SimpleScalar Alpha simulator, which is cycle accurate, but not timing accurate, we analyzed the number of timing errors occurring in the hardware model of a superscalar processor. We

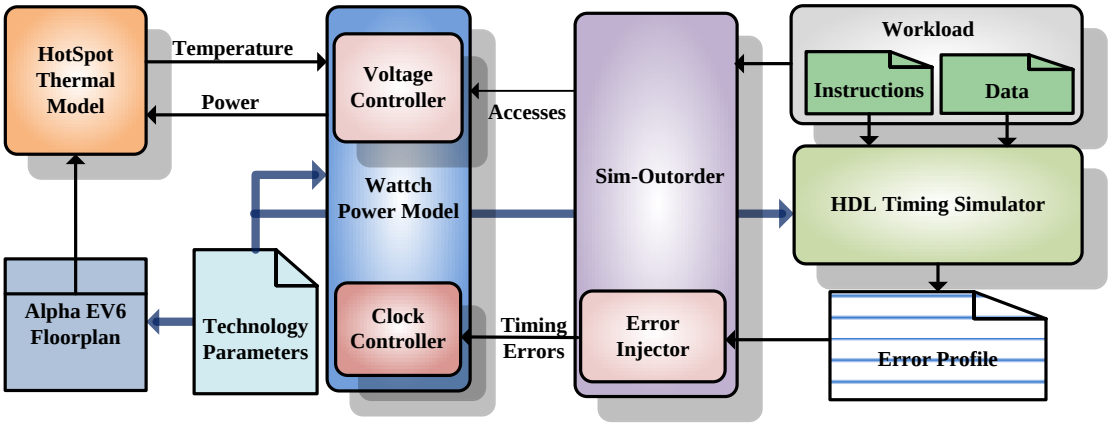

Fig. 6. Evaluation framework

Pipeline Stage	T_{PD} (ns)	T_{CD} (ns)	% Critical Registers	P_{leak} (mW)
Fetch	3.90	0.06	2.1	2.555
Decode	2.76	0.10	0	0.151
Rename	2.88	0.06	0	0.588
Issue	4.89	0.10	89.17	3.507
Execute	6.65	0.08	11.86	1.436
Memory	5.21	0.10	3.21	4.985
Commit	1.94	0.07	0	4.5

TABLE I
SYNTHESIS REPORT OF MAJOR PIPELINE STAGES

performed our study in a superscalar, dynamically scheduled pipeline similar in complexity to the Alpha 21264 [35]. We obtained the Illinois Verilog Model (IVM) from the University of Illinois website, which is a Verilog implementation of an Alpha microprocessor at the Register Transfer Level. The IVM Alpha processor executes a subset of the Alpha instruction set. However, the IVM processor is not fully synthesizable, and the synthesizable model does not support running SPEC 2000 benchmarks.

We designed the following experiment to evaluate the timing errors occurring at various overclocked frequencies. We synthesized individual pipeline stages using Synopsys design compiler. We used the $45nm$ OSU standard cell library for timing estimation [36]. There are altogether 12 pipeline stages in the processor. Table I reports the synthesis results for the major pipeline stages. In the IVM processor, the fetch stage, for instance, is divided into three stages. We report only the propagation delay, T_{PD} for the slowest among the three fetch stages. Similarly, we report the contamination delay, T_{CD}, for the shortest path across the three fetch stages. The timing values, reported in ns, are obtained from static timing analysis reports. However, the percentage of registers falling in the critical path and leakage power (P_{leak}) are reported for all three stages combined. In the table, we report the percentage of registers that have path terminating at them with delay values greater than or equal to $3.5ns$.

As explained in Section II-A, it is necessary to augment critical registers with error detection and recovery circuit, and also increase contamination delay of paths terminating at critical registers to a value greater than the desired extent of overclocking. Our simulator overclocks up to 40% of the worst case clock period. This requires the increase of contamination delay to over 40% the clock period. Using set minimum delay constraints in the Synopsys compiler, we increased the contamination delay to the required value. The overall increase in area for reliable overclocking was 3.5%.

Once we got the synthesized blocks for a particular stage, we replaced the RTL model for that block with the synthesized model. We also annotated timing information, extracted in standard delay format (SDF), on the blocks, so that we can run timing accurate simulations. We ran the instruction profile of various benchmarks obtained from the SimpleScalar simulator through the various stages. We used random data values for other inputs, filled the memory with random data. We measured error rate for various benchmark profiles.

Figure 7 shows the cumulative error rate for the IVM processor. We ran the experiment for 100000 cycles, and repeated the experiment with different sequences of 100,000 instructions for each benchmark. Average values are reported in the chart. We fixed the worst-case delay at $7ns$ to allow the maximum propagation delay of $6.5ns$ in the execute stage. We split 32 equal intervals from 7 to $3.5ns$ and measured error rate at each interval. We noticed around 89.17% of the paths fail in the issue stage at $3.5ns$, which causes a sudden rise in error rate, as observed in the Figure 7.

The error rate values we obtained from our hardware

simulation are incorporated in our functional simulator. While operating the simulator at higher frequencies, we use the error rates relatively. The leakage power, estimated at $105°C$ and $1V$ for IVM alpha processor at $45nm$, is used in Wattch power model to estimate leakage power at various temperatures.

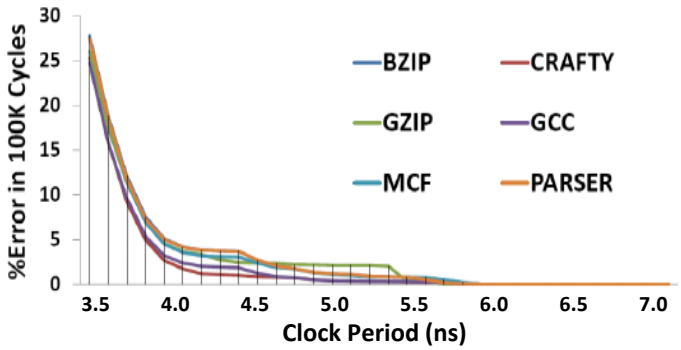

Fig. 7. Cumulative error rate at different clock periods for the IVM processor executing instructions from SPEC INT 2000 benchmarks

Parameter	Value
Fetch width	4 inst/cycle
Decode width	4 inst/cycle
Issue width	4 inst/cycle (ooo)
Commit width	4 inst/cycle
Functional units	4 INT ALUs 1 INT MUL/DIV 4 FP ALUs 1 FP MUL/DIV
L1 D-cache L1 I-cache L2 Unified	128K 512K 1024K
Technology node	$45nm$
Voltage	$1.25V$
Minimum frequency	$1024MHz$
Maximum frequency	$1862MHz$
No. of freq levels per voltage	32
Area	$10mm^2$
Temperature sampling	$1ms$
Freq sampling	$10\mu s$
Freq penalty	Single PLL: $10\mu s$ Dual PLL: $0\mu s$

TABLE II
SIMULATOR PARAMETERS

D. Area and Power Model for ROP

As explained in Section II-A, it is necessary to augment critical registers with error detection and recovery circuit, and also increase contamination delay of paths terminating at critical registers to a value greater than the desired extent of overclocking. Our simulator overclocks up to 45% of the worst-case clock period. This requires increasing contamination delay to over 45% the clock period. Using set minimum delay constraints in the Synopsys compiler, we increased the contamination delay to the required value. We adopt an overall increase in area of 3.5% for the additional circuitry from [7].

The power dissipation for the ROP was modeled based on our hardware implementation. We accounted for the additional area incurred for the local timing error detection and recovery circuits. For our model, we assume the increase in power dissipation to be directly proportional to the area increase we obtained. For each functional block in the processor the total power dissipation was increased by the fractional increase in overall area, obtained from our hardware experiments. The leakage power, estimated at $105°C$ and $1V$ for IVM alpha processor at $45nm$, is used in Wattch power model to estimate leakage power at various temperatures.

E. Simulation Parameters

Table II presents the simulation parameters. We evaluate the system temperature while running at $1.25V$. From Figure 2, we can see that clock period can be scaled only up to 50% of the original cycle time. We assume up to 45% overclocking. Table II provides the worst-case frequency and the maximum overclocked frequency we considered for our simulations. We perform a binary search between 32 frequency levels within the allowed range, based on error rate and also temperature, when employing thermal throttle. We assume the presence of two PLLs, so that there is no performance penalty involved, while switching between frequencies. If there is only one PLL, it takes up to $10\mu s$ to change from one frequency to another.

IV. On-chip temperature trends in reliably overclocked processors

We simulated six SPEC INT 2000 benchmarks to analyze the on-chip temperature trends in a reliably overclocked processor. The benchmarks reflect a broad spectrum of compute intensive workloads: gcc is a C compiler, $crafty$ is a chess program, mcf solves the minimum network flow problem, $parser$ involves natural language processing, $bzip2$ and $gzip$ are data compression utility applications.

We evaluate ROP performance with and without thermal throttling. Figures 8, 9, 10 and 11 compare the transient temperature trends and mean time to failure of a reliably overclocked processor with a non-overclocked processor for four of the benchmarks. Other two benchmarks have similar thermal characteristics. From the plots, we can clearly see that there is up to 15K difference between a reliably overclocked processor and a non-overclocked processor. Also, we see that the reliably overclocked processor reaches and exceeds 360K on executing around 3 million instructions. Based on the cooling solution used, the system will reach a steady state temperature and remain there. In our experiments, a non-overclocked processor settles at 347K for the same cooling solution. We start our experiments at a steady state temperature of 340K. This initial temperature is based on the assumption that the system has already performed certain operations, before it executes the benchmark of interest.

A. Frequency Based Thermal Throttling

When incorporating thermal throttle, we find that the temperature gets clamped at the desired choice of operating temperature. Since thermal sensor outputs are available once

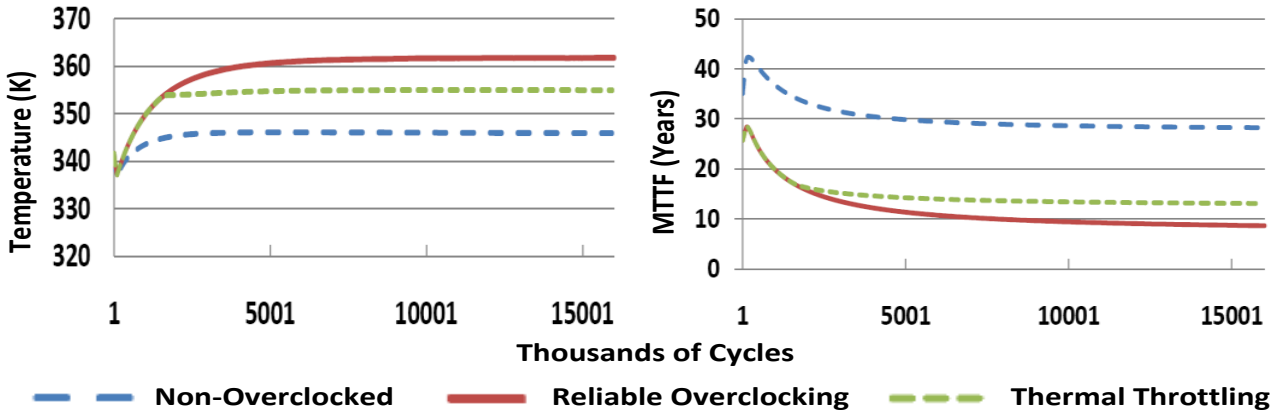

Fig. 8. $bzip2$ (a) On-chip Temperature Trend (b) MTTF Chart

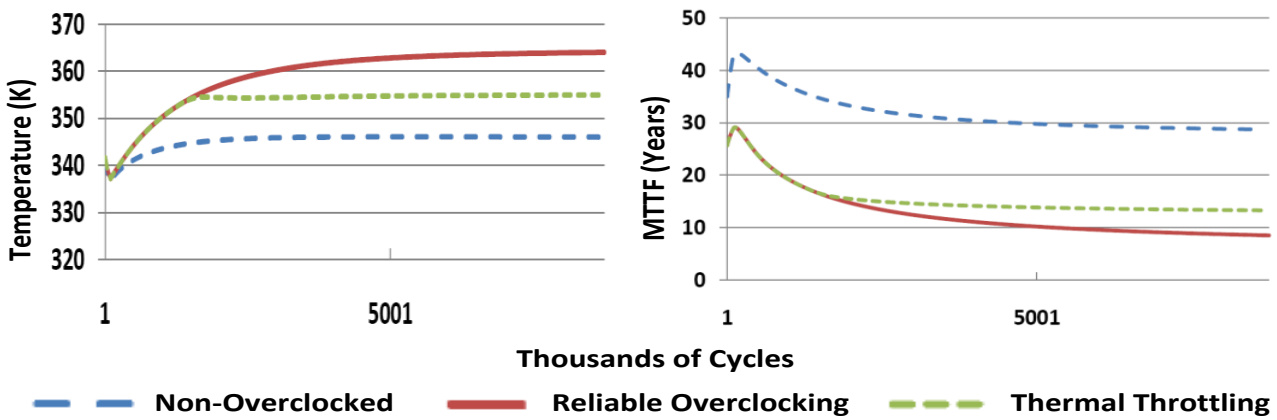

Fig. 9. $crafty$ (a) On-chip Temperature Trend (b) MTTF Chart

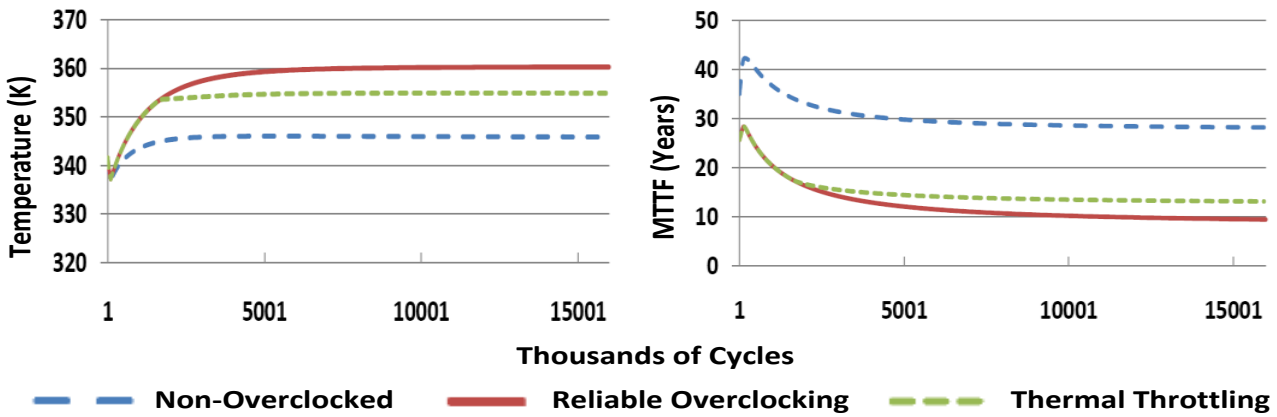

Fig. 10. $gzip$ (a) On-chip Temperature Trend (b) MTTF Chart

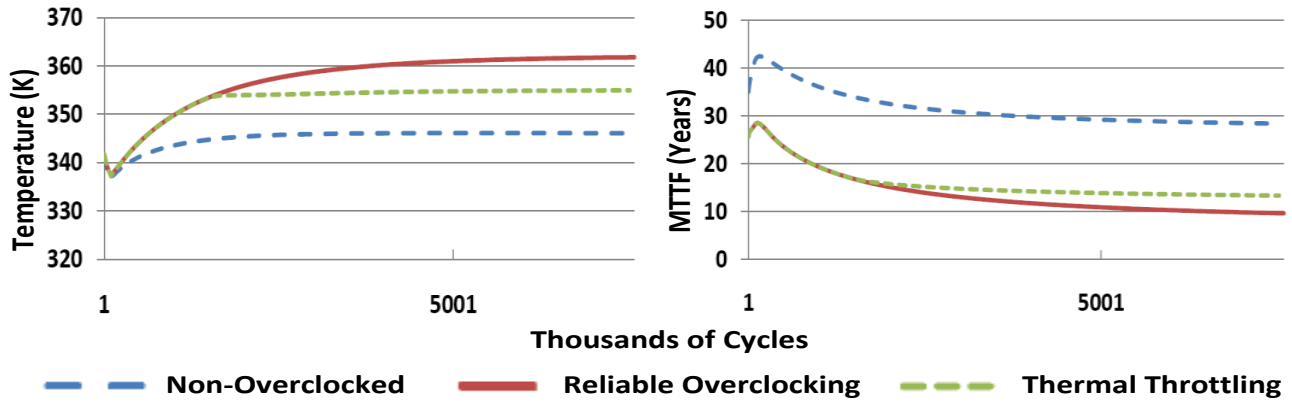

Fig. 11. mcf (a) On-chip Temperature Trend (b) MTTF Chart

every ms, it is good to choose a temperature 3K below the critical temperature, so that even if the system temperature overshoots before getting a thermal measurement, it will not exceed the critical temperature.

The relative speed-up for the six benchmarks, running 10^7 instructions is illustrated in Figure 12. Reliable overclocking, on an average, achieves 35% increase in performance over a non-overclocked system. When a thermal throttle is applied, the performance gain drops to 25%.

The MTTF values are obtained from the formulas mentioned in Section II-D. We calculate MTTF based on the on-chip temperature at that given instant of time. We obtain the proportionality constant for our calculations from the baseline MTTF at 337K [29]. We observe that a non-overclocked system has a longer lifetime, of about 30 years, as its on-chip temperature does not exceed 347K. However, a reliably overclocked system has a much shorter lifetime of about 9 years. Applying thermal throttling at about 355K increased the system lifetime to about 14 years. We understand from the figure that running a benchmark at lower temperatures over a long period of time improves the MTTF significantly. This motivates the need for having efficient dynamic thermal management techniques, alongside reliable overclocking, to achieve performance gain and reliability. Also, thermal management techniques alleviate the need for having an expensive cooling solution, making it cost effective to have high performance systems.

V. Conclusions

We have presented an initial study of the effects of reliably overclocked systems on on-chip temperatures. In addition, we also analyze the consequent effects on lifetime reliability of these systems. We considered a reliable overclocking framework and studied its thermal behavior compared to worst-case design. Our work in this paper is an initial exploration of dynamic thermal management in reliably overclocked systems. We are continuing this work by developing a powerful thermal management scheme that enhances performance as much as possible while operating well within the thermal limits, guaranteeing an extended system lifetime. The results we have

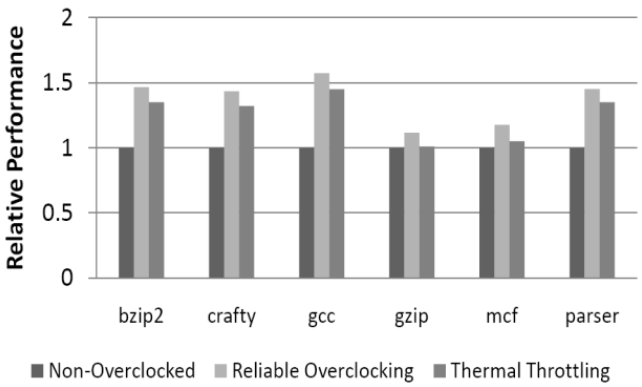

Fig. 12. Relative performance for SPEC INT 2000 benchmarks

obtained at this juncture are very promising, opening up many different directions for the near future.

We would like to further this work by implementing it on a hardware platform such as FPGA, by tracking temperature online through thermal sensors. We also plan to test our scheme with an ASIC model. As an extension to this work, we are planning to combine this technique with the existing dynamic voltage and frequency scaling (DVFS) and unify them within a common framework for better power saving. Finally, this work also opens up a new direction of managing power in chip multiprocessors, where our technique has the potential to allow fine grained and more accurate power management across the cores.

Acknowledgment

The research reported in this paper is partially supported by NSF grant number 0311061 and the Jerry R. Junkins Endowment at Iowa State University.

References

[1] V. Subramanian, P. K. Ramesh, and A. K. Somani, "Managing the Impact of On-Chip Temperature on the Lifetime Reliability of Reliably Overclocked Systems," in *Second International Conference on Dependability-DEPEND'09, June 18-23*, 2009.

[2] *http://www.crn.com/hardware/212101254*.

[3] R. McGowen, C. Poirier, C. Bostak, J. Ignowski, M. Millican, W. Parks, and S. Naffziger, "Power and temperature control on a 90-nm Itanium family processor," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 229–237, 2006.

[4] E. Rotem, A. Naveh, M. Moffie, and A. Mendelson, "Analysis of thermal monitor features of the intel pentium m processor," in *TACS Workshop at ISCA-31*, 2004.

[5] "AMD PowerNow! Technology," http://www.amd.com/epd/processors/6.32bitproc/8.amdk6fami/x24267/24267a.pdf, Date Accessed: March 31, 2009.

[6] D. Ernst, N. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, *et al.*, "Razor: A low-power pipeline based on circuit-level timing speculation," in *MICRO-36, 36th Annual IEEE/ACM International Symposium on Microarchitecture*, 2003, pp. 7–18.

[7] V. Subramanian, M. Bezdek, N. Avirneni, and A. Somani, "Superscalar processor performance enhancement through reliable dynamic clock frequency tuning," in *IEEE/IFIP International conference on Dependable Systems and Networks*, 2007, pp. 196–205.

[8] D. Burger and T. Austin, "The SimpleScalar tool set, version 2.0," *ACM SIGARCH Computer Architecture News*, vol. 25, no. 3, pp. 13–25, 1997.

[9] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis and optimizations," in *Proceedings of the 27th Annual International Symposium on Computer architecture*. ACM New York, NY, USA, 2000, pp. 83–94.

[10] W. Huang, K. Sankaranarayanan, R. Ribando, M. Stan, and K. Skadron, "An improved block-based thermal model in HotSpot 4.0 with granularity considerations," in *Proceedings of the Workshop on Duplicating, Deconstructing, and Debunking*, 2007.

[11] S. . B. Suite, http://www.spec.org/cpu2000/.

[12] A. Uht, "Uniprocessor performance enhancement through adaptive clock frequency control," *IEEE Transactions on Computers*, vol. 54, no. 2, pp. 132–140, 2005.

[13] A.K.Uht, "Achieving typical delays in synchronous systems via timing error toleration," in *Technical report 032000-0100*, Dept. of Electrical and Computer Eng., Univ. of Rhode Island, Kingston, 2000.

[14] T. Sato and I. Arita, "Constructive timing violation for improving energy efficiency," in *Compilers and operating systems for low power*, Kluwer Academic Publishers, 2003.

[15] X-Vera, O.Unsal, and A. Gonzalez, "X-pipe: An adaptive resilient microarchitecture for parameter variations," In Workshop on Architectural Support for GigascaleIntegration, 2006.

[16] T. Austin, V. Bertacco, D. Blaauw, and T. Mudge, "Opportunities and challenges for better than worst-case design," in *ASP-DAC '05: Proceedings of the 2005 Asia and South Pacific Design Automation Conference*. New York, NY, USA: ACM, 2005, pp. 2–7.

[17] G. Memik, M. Chowdhury, A. Mallik, and Y. Ismail, "Engineering over-clocking: reliability-performance trade-offs for high-performance register files," in *Dependable Systems and Networks, 2005. DSN 2005. Proceedings. International Conference on*, June-1 July 2005, pp. 770–779.

[18] V. Subramanian and A. K. Somani, "Conjoined Pipeline: A Fault-Tolerant High Performance Microarchitecture," in *Pacific Rim International Symposium on Dependable Computing, , Taipei, Taiwan, Dec*, 2008.

[19] N. D. Avirneni, V. Subramanian, and A. K. Somani, "Low Overhead Soft Error Mitigation Techniques for High-Performance and Aggressive Systems," in *IEEE/IFIP Dependable Systems and Networks*, 2009.

[20] F. M. J. Renau, "Effective optimistic checker tandem core design through architectural pruning," In International Symposium on Microarchitecture, 2007.

[21] T. M. Austin, "Diva: a reliable substrate for deep submicron microarchitecture design," in *MICRO 32: Proceedings of the 32nd annual ACM/IEEE international symposium on Microarchitecture*. Washington, DC, USA: IEEE Computer Society, 1999, pp. 196–207.

[22] S. Kim and A. K. Somani, "Ssd: An affordable fault tolerant architecture for superscalar processors," in *Pacific Rim Dependable Computing Conference*, December 2001, pp. 27–34.

[23] A. Tiwari, S. R. Sarangi, and J. Torrellas, "Recycle: Pipeline adaptation to tolerate process variation," 34th International Symposium on Computer Architecture (ISCA), 2007.

[24] S. Sarangi, B. Greskamp, A. Tiwari, and J. Torrellas, "Eval: Utilizing processors with variation-induced timing errors," In International Symposium on Microarchitecture, 2008, pp. 423–434.

[25] S. Sarangi, B. Greskamp, and J. Torrellas, "A model for timing errors in processors with parameter variation," in *Quality Electronic Design, 2007. ISQED '07. 8th International Symposium on*, March 2007, pp. 647–654.

[26] B. Greskamp, L. Wan, U. Karpuzcu, J. Cook, J. Torrellas, D. Chen, and C. Zilles, "Blueshift: Designing processors for timing speculation from the ground up," International Symposium on High Performance Computer Architecture, 2009, pp. 213–224.

[27] *PowerPC 750GX Dynamic Power-Performance Scaling*. Version 1.0, January 29, 2009: IBM Systems and Technology Group. [Online]. Available: http://www.ibm.com

[28] J. H. et al, *A Robust Physical and Predictive Model for Deep-Submicrometer MOS Circuit Simulation*. Proceedings of the IEEE Custom Integrated Circuits Conference, pp.14.2.1-4, May 1993.

[29] J. Srinivasan, S. Adve, P. Bose, and J. Rivers, "Lifetime reliability: Toward an architectural solution," *IEEE Micro*, vol. 25, no. 3, pp. 70–80, 2005.

[30] K. Skadron, M. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware microarchitecture," in *Proceedings of the International Symposium on ComputerArchitecture*, 2003, pp. 2–13.

[31] S. Heo, K. Barr, and K. Asanovic, "Reducing power density through activity migration," in *ISLPED'03, Proceedings of the 2003 International Symposium on Low Power Electronics and Design*, 2003, pp. 217–222.

[32] P. K. Ramesh, V. Subramanian, and A. K. Somani, "Thermal Management in Reliably Overclocked Systems," in *IEEE Workshop on Silicon Errors in Logic - System Effects, Stanford, CA, Mar*, 2009.

[33] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, *The Case for Lifetime Reliability-Aware Microprocessors*. The 31st International Symposium on Computer Architecture (ISCA-04), June 2004.

[34] J. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital integrated circuits*. Prentice Hall Englewood Cliffs, New Jersey, 2002.

[35] N. Wang, J. Quek, T. Rafacz, and S. Patel, "Characterizing the effects of transient faults on a high-performance processor pipeline," in *International Conference on Dependable Systems and Networks*, 2004, pp. 61–70.

[36] J. Stine, I. Castellanos, M. Wood, J. Henson, F. Love, W. Davis, P. Franzon, M. Bucher, S. Basavarajaiah, J. Oh, *et al.*, "FreePDK: An Open-Source Variation-Aware Design Kit," in *Proceedings of the 2007 IEEE International Conference on Microelectronic Systems Education*. IEEE Computer Society Washington, DC, USA, 2007, pp. 173–174.

Biodiversity Information Systems Evolution:

The MABIS model to gather several communities on an adaptable environment

Didier Sébastien, Noël Conruyt,
Rémy Courdier, Nicolas Sébastien
IREMIA, LIM-EA2525,
Université de la Réunion
Saint-Denis, France
didier.sebastien/noel.conruyt/remy.courdier/
nicolas.sebastien@univ-reunion.fr

Tullio Tanzi
Département TSI - LTCI UMR 5141 - CNRS
Institut TELECOM - TELECOM ParisTech
Paris, France
tullio.tanzi@telecom-paristech.fr

***Abstract*—The computerization of scientific data treatment and the relatively recent awareness of the fragility of the natural world environment (Rio Conference "Earth Summit" in 1992) have led to the proliferation of Information Systems dedicated to biodiversity. Given the various data that they contain, they are complex applications, focused on the needs of one type of environmental scientist, closed to amateur's contributions and unable to support ethological information. Moreover, the data contained in these systems are hardly accessible to non-specialists like the general public or decision-makers.**

At the same time, new communication protocols, like webservices ensure a better sharing of information between applications; while immersive representations of three-dimensional virtual worlds, also known as metaverses, allow a more natural assimilation of information. By putting users in a reality reproduction built from information systems, all entities can be represented in a consistent virtual environment [15]. But sharing and turning Biodiversity Information System's data in a coherent metaverse is not a trivial process. It relies on an adapted architecture and the availability of specific metadata, as several virtual worlds representing several levels of details can be generated.

This paper is a contribution to the enhancement of Biodiversity Information Systems in order to make them more adaptable, usable, and representable for several kinds of users: different types of specialists, amateurs, and the general public. The MABIS modular architecture is presented as an open, efficient and evolutionary model for structuring Biodiversity Information System. The advantages of its architecture allow to ease its development, store ethological data, manage, share and represent complex entities and metadata while ensuring their authentication through an evaluation process.

Keywords: Biodiversity Information System, BIS, MABIS, behavior, immersion, metaverse, data evaluation, webservice

I. INTRODUCTION

The important loss of world biodiversity has led Eco-Informatics experts [18] to develop specific Environmental Management Information Systems (EMIS) called Biodiversity Information Systems (BIS). The objective is to help communities of practice in Biology (thematic experts, also called "thematicians") who work on this phenomenon in their everyday tasks: produce inventories of species, obtain a better visibility on biodiversities' studies, plan coordinated actions between institutions and communicate their results. In this context, Information Systems (IS) are more and more used to manage biodiversity data [6].

Most BIS focus on the management of taxa and specimens [9], but the available functions and uses are directed to fit the needs of a specific category of scientific specialists. For the same substrate (environmental information), it is possible to define three different types of users and needs: researchers, managers, and curators. Each of these professions works with different objectives in the environmental field and thus has different needs in terms of BIS functions (see Figure 1).

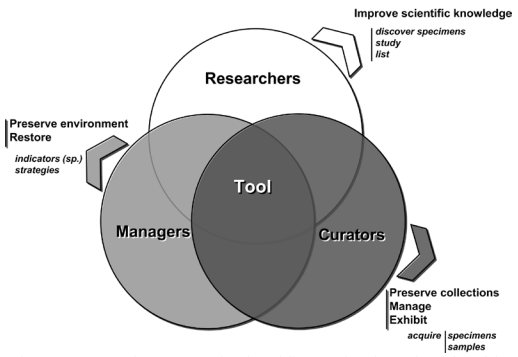

Figure 1. Main types of scientific professions involved in environmental data management

The researchers' objective is to make scientific progress by accumulating new knowledge. They accomplish it mainly by discovering new specimens, analyzing them in laboratories (*ex situ*), and studying them by experimentation (*in situ*). On the global scale, the objective is to understand the diversity and evolution of the life through time, space and interactions, as well as the mechanisms at the origin of this diversity.

The objective of environment managers is to discover ways to preserve and restore the ecosystems for which they are responsible. To achieve these goals, they need to uncover specific indicators that characterize the environment and they

must accurately adhere to the policies of environmental restoration they have developed.

The curators' aim is to preserve and manage collections. Fulfilling this task requires acquiring new specimens, sorting and arranging them in order to preserve the taxonomists' primary tool and biodiversity hot spot, but also setting up expositions destined for the general public.

As we can see, in these three scientific vocations specimens are manipulated, but apart from sharing a metadata basis, there is a need to associate different types of entities (documents, maps, collections) with different levels of granularity. That is why, even if their field of work is the same, researchers, environment managers and curators will use a specific BIS to manage their data. This however, implies *lost time* in repeated metadata entries in the absence of data sharing and communication between the communities that work on the same set of taxa.

Moreover, *amateurs are not allowed to contribute* on the system by adding their own data in order to share them with the scientific community. Indeed, the biodiversity field attracts a parallel community of non-specialists which gathers important sets of data. Most of time, these data are rejected by BIS only because they were not provided by experts. However, to profit from these data, it would be useful to let amateurs have a controlled access on the system.

Another common limitation of BIS is they *do not support ethological information* as a specific data type. Because ethological studies are not considered as a specific experimentation, specimen's behavior is not recognized as an independent entity on the system, and researchers studying them have no access to specialized and structured forms, like ethograms, to enter this type of information.

Lastly, biodiversity information systems contain a huge quantity and variety of data that lead to important *assimilation's difficulties*, especially for non-specialists of the thematic. Thus, it is often difficult for thematicians to communicate their results to the general public or the decision makers through the classic Graphic User Interface (GUI) of BIS.

Therefore, we propose an open architecture that tries to find a place for each type of users, from the experts to the general public. The aim is to make information systems more adaptable, usable, and representable for several kinds of users; without losing the scientific credibility.

II. THE MABIS MODULAR ACHITECTURE

A. General architecture

In order to share common data and knowledge between distinct BIS, it is possible to create a complete set of web services, which translates into a great amount of development for each platform. Another possibility would be to set up an adaptable BIS, corresponding to the needs of each type of user, without imposing a complex Graphic User Interface. Relying on this approach, we propose the Modular Architecture for Biodiversity Information System (MABIS). This model uses several Web applications that constitute modules interconnected together by webservices (Figure 2).

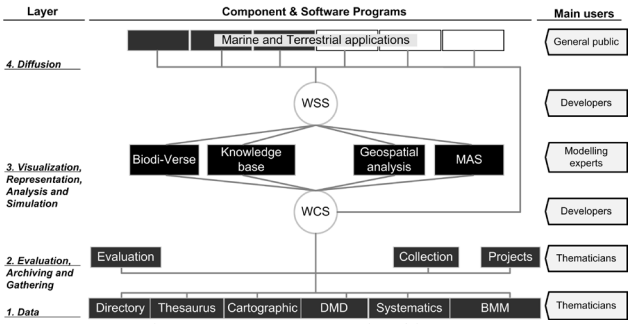

Figure 2. MABIS general architecture

These Web applications are structured in four layers:

- The first layer is dedicated to the management of the main BIS entities, i.e. the directory of participants to the project, the thesaurus for explaining the meaning of terms, the Documentary Multimedia Database (DMD) [14] for the studied objects, the systematics module for storing, organizing and presenting specimens and taxa, the cartographic module for georeferencing them, and the Behavior Management Module (BMM) to register ethological data,
- The second layer provides tools to authenticate the information stored in the first layer (evaluation module), to allow curators to archive collections' data about specimens, and to manage the several types of entities from the first layer in the frame of contextual environmental projects, for example to provide managers the means to monitor their protected areas (follow-up module),
- The third layer provides the software tools to visualize, represent, analyze and treat the data from the first two layers. For instance, the Biodi-Verse software is dedicated to the generation of immersive views for selected data, whereas the Multi-Agent Simulation (MAS) software offers simulation's possibilities with the codified information on specimens [2],
- The fourth layer is dedicated to vulgarization. It hosts Web portals relying on data supported by the other layers.

Figure 3. Screenshot of the connexions between modules' GUIs

More than the advantage of facilitating its development, thanks to its modular architecture, MABIS let each type of user focuses on his immediate task: each module constitutes

an entry point (Figure 3) corresponding to precise expectancies and functionalities. Furthermore, to ensure exchanges between the MABIS architecture and other BIS, two gates allow remote controls on the main layers: the Web Components Services (WCS) gate opens on functionalities offered by the two first layers, whereas the Web Software Services (WSS) gate provides a convenient access to the third layer.

B. *The modules*

The modules constituting the IS are applications dedicated to the management of a limited set of entities' types. All the information related to an entity (for example, a document) is gathered on a structured card which can be enhanced with elements from another entity (Figure 4).

Each module has three functioning modes: main, deported, and remote mode. In *main mode*, users work directly in the environment of the Web application dedicated to the management of the entity he is focusing on. The *deported mode* is used to provide popup windows and inclusions offering synthetic graphical information's and functionalities from a module in another one used in main mode. The *remote mode* means that the Web application is used through its webservices, in order to fully integrate its data and functionalities in another module's GUI.

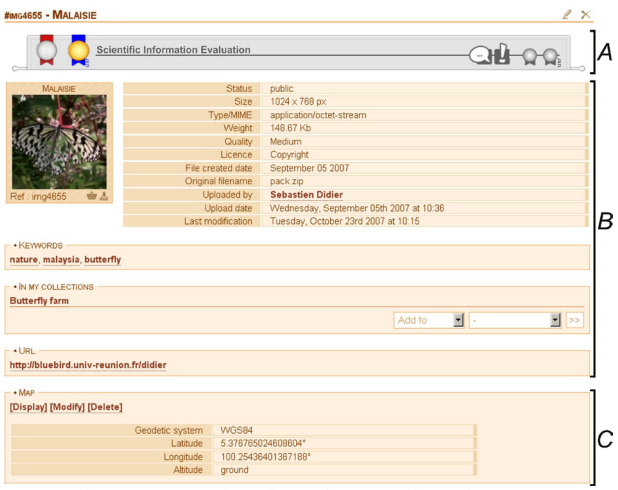

Figure 4. Screenshot of a card representing a document
A: deported mode from the evaluation module
B: main part of the card in the main mode of the multimedia database
C: data provided by the remote mode of the cartographic module

Moreover, a distinction can be made between modules, taking into account their complexity. The Web applications of the first and second layers are dedicated to the management of entities. They can be fully used through Web browsers in the three modes and constitute WCS. By contrast, the Web applications of the third layer, provide complex treatment tools on primary data of the first layer, so rely on a main mode requiring a typical installation on computers. Thus, this mode is mainly not working like a SaaS (Software as a Service) application. Only deported and remote modes can be used through Web Browser. These modules constitute WSS.

In order to make easier the assimilation of the Web Components Services, which represent the common entry points of new primary data, we have standardized them to keep the same logic of entities' management. Indeed each entity's type has to support common categories to guaranty a good communication between contributors. The simple template represented on Figure 5, established through a human centered development [14], has led the general organization of all the WCS deployed in MABIS.

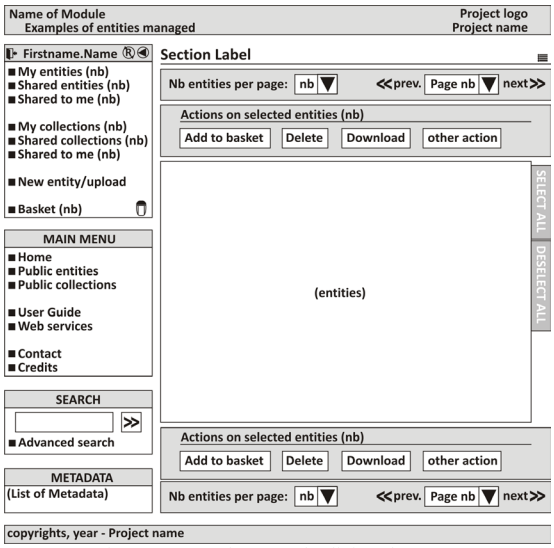

Figure 5. Template standardizing the WCS

An instance of this template can be seen on Figure 9.

III. THE EVALUATION OF SCIENTIFIC INFORMATION

The biodiversity field attracts a community of enthusiasts, non-expert of the thematic: amateurs. They gather a huge quantity of data that are often not accepted on BIS because the providers are not specialists. A solution to this loss of information is to support a scientific data validation policy, which is often made by an administrator in Biodiversity Information Systems. However, because the quantity of data expert-administrators have to validate is colossal, they cannot afford to treat amateurs' resources.

In this context, we propose a Scientific Information Evaluator (SIE) model. This model relies on the MABIS architecture and comes into the form of a SIE module that manages the evaluation of all entities' cards distributed in the IS. Our aim is not to provide a methodology to analyze data, for instance by error elimination [12], but to deliver a tool that follows the enhancement of a card until it reaches the best acknowledgement level. This module offers several ways to authenticate the scientific aspects of data, to share the evaluation's work, to simplify the communication between experts, and to ease the identification of the evaluation's state by end-users. The enhancement process can be represented as in Figure 6.

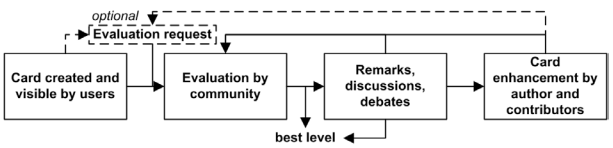

Figure 6. Evaluation of scientific information in MABIS model

Contrary to most validation processes that result in a binary answer (validated/not validated), evaluations result in certificates associated to a precise time. Each certificate corresponds to a level of trustworthy that gives a more precise idea of the data condition (Figure 7). The certificates can be considered as a succession of steps toward the validation level. In order to reduce the work of data evaluation by experts, two evaluation systems are offered.

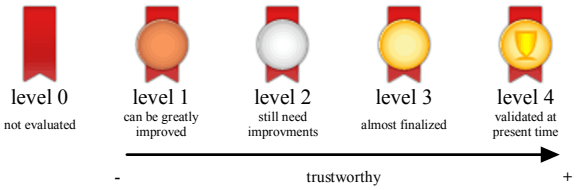

Figure 7. Levels used for the evaluation

The *authoritative certification* represents the evaluation of the current version of the entity's card made by identified specialists of the thematic on the BIS. Depending on their recognized specialization in a discipline, a limited set of experts acquire from the administrator the possibility to deliver certificates that represent their approval to the data. This ability to deliver up to a defined level of certification is entity-specific, and for the taxa, taxa-specific. So the SIE stores a list of privileges that can be associated to profiles of users registered in the directory. The propagation of this certification is regulated by a cooptation system that let specialists share their privileges to other users up to their own level. Thus the administration and the evaluation work are shared. This evaluation based on hierarchy is summed up by a simple icon that represents the lowest certification given by the highest specialist. Experts' debate, i.e. when two or more specialists at the same highest level deliver different certificates for the same card, is automatically detected and notified by the sign "!". In this case, users are warned that it could be necessary to read the comments and evaluate themselves the card.

The *community certification* is the evaluation of the card made by all identified users on the BIS. Because everyone can participate and deliver the certificate they consider deserved, this evaluation is less specialized than the precedent one; however, a certificate coming under an important participation of users gives a good basis of evaluation if experts have no time to consider the card. This notation based on folksonomy is summed up by an icon that shows the average level of certification given by voters.

These two certification modes are easy to use (a simple button) and can be used as criteria, in a combined way or not, to sort and research data. Thus, it is easy to work only with, for instance, data defined as relevant by experts.

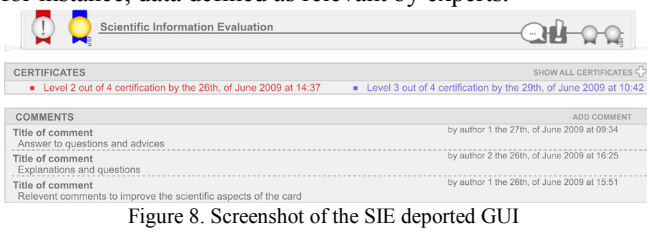

Figure 8. Screenshot of the SIE deported GUI

The Scientific Information Evaluator's deported mode (Figure 8) adds three inclusions in the graphical interface of cards, to:

- indicates the level of validation, which represents the level of approval of the information. Each level is represented by an icon that sums-up the participative evaluation of a data.
- presents detailed information and an historic about the evaluation of the card (all marks from all voters),
- offers a simple thread that permits authors and evaluators to discuss about the card.

The SIE's main mode (Figure 9) is the Web application that gathers all information about evaluation of entities. It allows to:

- present general metadata about entities' evaluation on the BIS (most discussed cards, most heterogeneous evaluation per module, *etc.*).
- list all evaluations' requests and focus on those concerning the authenticated user.
- show the logged in user his own evaluation requests, delivered certificates, discussions' threads (new messages since last connection), and the current evolution of his cards.
- manage the sharing of privileges concerning the authorized level of certification.

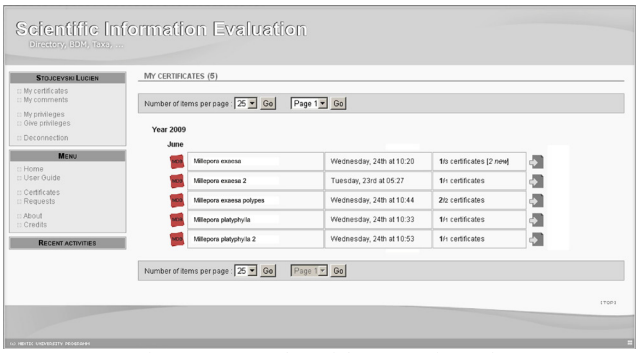

Figure 9. Screenshot of the SIE main mode

The remote mode of the SIE is of course not related to a specific GUI, but it is often used in other modules, for instance to sort entities by certification level in other Web applications' main mode.

IV. THE MANAGEMENT OF ETHOLOGICAL INFORMATION

A. *States and actions, a difference between specialists' approach and system's structure*

Ethology is the science that studies the behavior of living species. Biologists that analyze the behavior of animals (also called "ethologists") focus their researches on the establishment of lists of states and actions, the determination of their occurence and linking. By observing several individuals of a taxon, they identify their way of life, gaits, interactions between them and other species, impact on their ecosystem, etc. These observations are usually made during short analyses' sessions, on a limited number of specimens; so a lot of data is produced in a short duration, often associated to a precise scale of time (minutes or seconds). This kind of scale is hardly supported by common BIS, on

which temporal data associated to a biodiversity project (like a monitoring) uses larger time's scales (days).

Independently of the structural aspect of BIS that has to evolve to handle the information associated to precise scales of time, the user interface has to be enhanced in order to facilitate the data's entry process. That is why the best way to integrate this dimension without compromising the human-computer interaction is, in the frame of our architecture, to add a new WCS: the Behavior Management Module (BMM). As we have said, ethologists focus on states and particularly actions (movements) performed by the studied species. Indeed states and actions are linked: particular characteristics of internal (not visible) and external (visible) state lead to specific actions. Without a sustained analysis of a specimen, it is hard to evaluate its state (for instance, his hunger): researchers have to determine it from the actions that will be induced. So even if a state leads to specific actions, for animals, ethologists often deduce the precedent state from the following action. That is why, to follow their methodology of work and processes, our module focuses more on actions. However, by developing tools permitting biologists to gather and represent dynamical data through a formalized approach, this work also contributes to initiate a discussion with experts to enhance their methodology of study.

B. Implementation in the MABIS architecture

In the frame of our architecture, entities from the WCS are intended to be used by the WSS as primary resources. Thus, it is interesting to take into account the use that will be done with the ethological data since the conception of the application. Specimen's behavior information, by facilitating the extraction of rules leading agents used in MAS, constitutes relevant information to build simulations. Moreover, the description of actions stored in the behavior module can be useful to generate immersive representations of data. That is why, in the implementation of this module, we have anticipated the use that will be made in terms of data treatments and representations.

	Behavior	
	State	***Action***
Text	body description	description of body movements
Representation	photo/icon/model of different poses	animation between static poses
MAS	list of relevant factors and way to determine their value	Rules of variation of the values

TABLE I. DESCRIPTORS FOR STATES AND ACTIONS

Current BIS are able to handle temporal data associated to specimens as a succession of observations describing its general state [2]. If this information could be sufficient in some cases, for instance phenological studies on plants, it is not in many others, like behavioral studies on animals. Finally, only little information can be automatically extracted from Systematics module to generate rules necessary to build Multi-Agent Simulation, as the observations are consigned in generic text fields. So there is a strong need to formalize spatio-temporal metadata about specimen's evolution and behavior. However, there are several ways to manage behavioral information in BIS. For instance, it is possible to consider either only one default state with actions as variations to this referential, or several states associated to actions that make the links between them. In order to clearly separate actions and states, we chose the second possibility structured through three approaches: textual description, visual representation, and formalized rules describing specimens' behavior as agents in MAS (TABLE I.). The objective is to establish a consistent timeline (Figure 10) made by states and actions, linked to the Systematics module, and ready to provide a formalized support to the immersive representation and ecosystems' simulation.

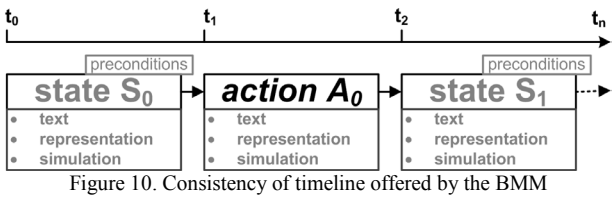

Figure 10. Consistency of timeline offered by the BMM

The way specimens' states are described also has to change for a more generic aspect. This directly impacts on the general Systematics module's structure. Instead of describing several times the same state of a specimen as several observations, it is possible to define a set of generic states associated to a taxonomic level. Doing this way, an observation on a specimen becomes the junction between a generic action of the taxon, and a specialized scientific annotation (Figure 11). The process is, of course, the same for states. Indeed the introduction of a BMM in BIS, and its strong relation with the module in charge of taxa and specimens' management lead to several evolutions of the Biodiversity module. The main improvements are explained in parallel of the descriptions of the three BMM's sides, but first, we have to take over global aspects of this tool.

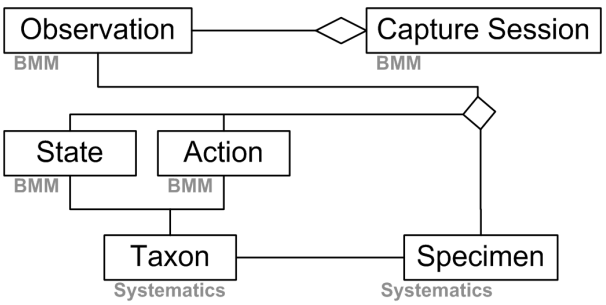

Figure 11. BMM entities' integration in MABIS

The BMM aims to manage specimen's behavior in two ways. The first one is to store labeled descriptions of actions and states achieved by specimens in the frame of a behavior process to achieve a goal. Ontologies are constituted by linking actions to a taxonomic rank in order to specialize or generalize it. For instance, the action "eat" can be defined in general for mammalians, and a second occurrence can be added to specialize the action for the Felidae family. Lists of different actions focusing on the same taxa can be gathered on a single view in order to constitute ethograms.

The second one is to store lists of actions related to specimens to constitute spatio-temporal capture sessions that

focus specimens. Indeed ethologists mainly study animal behavior through three methods:

- Focus sampling with continuous recording, which focuses and list all actions achieved by a specific specimen during a very limited time.
- Scan sampling, which point out all relevant actions achieved by several specimens.
- *Ad libitum* sampling, which means that the observer records as much as he can of whatever he can see. This method is generally used to get an overview of the specimen's group.

In all cases, the user has to associate or describe actions made by one specimen or more. Digitally speaking, it is just a matter of presentation, by action or by specimen. After entering the data on the system, it is easy to sort them by using the BMM.

So, in terms of use cases, the BMM answers to several users needs. Firstly, it will help ethologists to provide standardized, easily interoperable by computational request, action cards. Secondly, they will easily constitute and share sampling sessions on the Web. Thirdly, it will allow them to easily communicate with the general public their observations by showing immersive reconstitutions through immersive representations. Fourthly, simulation experts, will access to the module as a warehouse of generic and specific models for their MAS, knowing they could possibly represent their results in an immersive way. Fifthly, the collection of 3D models, animated or not, gathered in the frame of the BMM will constitute a warehouse [5] of species' shapes and movements for computer graphics experts.

As we have explained, states are often deducted from the following actions, so action represents the main point for ethologists. In this frame, we will now focus on how the "action" entity can be integrated through the three approaches in the same module.

C. Three representations for one action

To build a reliable and generic action's description form, that allows a relevant generation of action's description cards, we have to consider the definition of "behavior". For J.B. Watson, the behavior is "the whole of the objectively observable reactions that an organization generally equipped with a nervous system carries out in answer to stimulations of the medium, themselves objectively observable". Note that this definition is particularly opened, and underlines following key points. Behavior:

- refers to any specimen, whatever its phylum is (not restricted to animals),
- considers the observation's environment,
- focuses on objective actions.

Regarding these clues, and usual form's fields, we tried to build a generic action's description form that will gather main data (Figure 12). This general part contains textual and numerical data that are used by several users (ethologists, simulation experts, computer graphics experts). These fields allow multicriteria searches to retrieve specific actions and are highly used by the simulation and the representation engines (especially the duration and the periodicity fields).

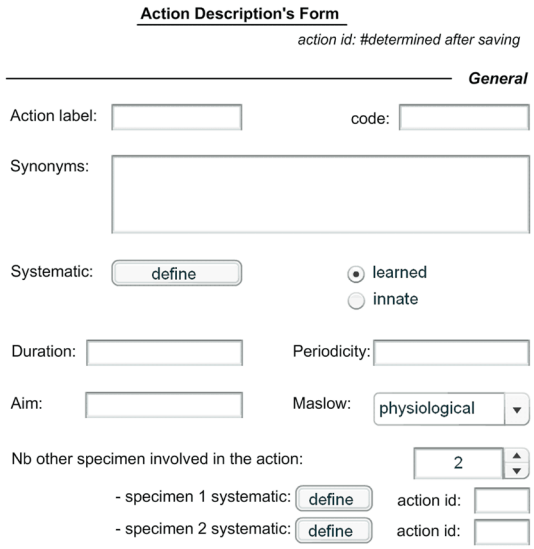

Figure 12. General description fields of the behavior form

In the frame of a BIS's module, the aim is to formalize the descriptions in order to meet both user's expectations and system's requirements, for instance, to avoid future plain text searches. The consistency of the integration of the behavior module in the IS relies on the links each entity (action, state, and sampling session) for each approach (text, simulation, representation) develops with other modules. Given the nature of the entities and their metadata, the systematic and the cartographic modules are particularly important in their description cards.

1) Descriptive aspect

The textual description (Figure 13) of specimen's behavior constitutes a classic approach of an ethological study. It is dedicated to specialists in Ethology, as the technical vocabulary they use is sometimes unreachable for the general public. That is why visual representation is so important. The general context field describes the conditions that must be gathered, for the specimen and the environment, to produce the action. It is described in a more formal way in the simulation aspect's part.

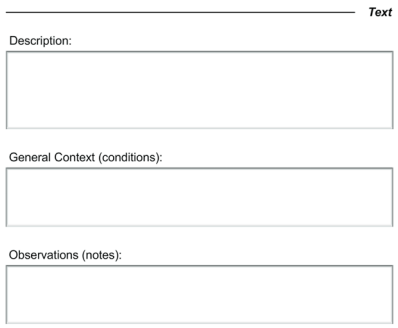

Figure 13. Textual description fields of the behavior's form

The gathering of main research fields in the general form permitted to simplify the textual description ones. That is why this part allows specialists to work in the way they are used with.

2) Representation aspect

This part collects the files made by ethologists and computer graphics experts to graphically represent the actions. The system accepts as inputs several files corresponding to several representations. Traditional graphical representations, in two dimensions, is relevant in the frame of the IS because it allows thematicians to gather all their data on the same system. Indeed, with the affordability of digital camera, thematicians produce more and more videos and photos to illustrate their researches. Furthermore, immersive environment are able to display this information on 3D surfaces as animated textures when tridimensional meshes are not available. That is why we chose to support this usual approach on the graphical action's description card's part (Figure 14). However, as we also aim at representing immersive representations, the application gathers all files needed to prepare this process.

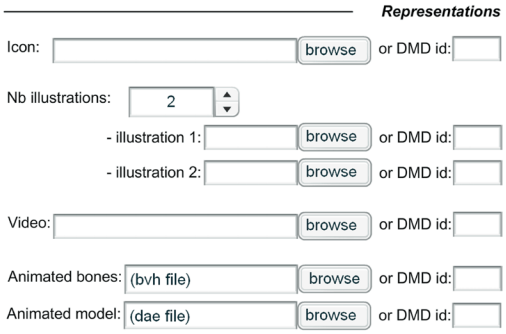

Figure 14. Form dedicated to graphical representations

Thus, several graphic-representation files are supported:

- The icon is an illustration created by the user to represent a specific action of a species. It can be used either in the IS to graphically sum-up a sampling session, or by the multi-agent simulation engine for its usual 2D visualization, or by the 3D engine as a texture.
- The photos and videos are multimedia documents representing actions, linked to its card. Because biologists usually take several photos and sketches to illustrate the same action, it is important that the system allows them to gather all their pictures by handling multiple file uploading.
- The "Animation bones" field refers to a tridimensional spatio-temporalized description of animation, in the BVH (BioVision Hierarchy) format, which is very used by motion-capture ("mocap") systems. Indeed, an animation description file can be built with several methods (handmade by a computer graphic expert, or with an automatic acquisition system). The description of action can then be applied to the 3D model of the specimen in order to make it reproduce the movement. This step requires a computer graphic expert to help the thematicians.
- Because it is not always possible to apply an animation to the 3D shape linked with the specimen in the Biodiversity module, a possibility is given to directly upload an animated model that shows the 3D reproduction of an action for the specimen. In this case, the COLLADA format, which stands for "Collaborative Design Activity", is favored, as it is an open and compatible file.

All files linked to this part are stored and managed by the documentary multimedia database of the IS, and shared by webservices in a transparent way for the user.

3) Simulation Aspect

The Multi-Agent Simulation (MAS) aspect of the Behavior Management Module aims at gathering information concerning taxa behavior for multi-agent simulation. Toward the generation of metaverses, we use the multi-agent paradigm as presented by Ferber [3] to define the different components that take parts in simulation:

- Agents are used to represent specimens. Indeed, agents are autonomous entities that have their own partial perception of the environment they live in. They also interact with other agents and with the environment.
- Environment is used to describe the landscape of the simulation and the global evolution rule of the ecosystem. In multi-agent simulation, the environment is the world agents evolve in. It can adopt many topologies: spatial or non spatial, with metrics, etc. The environment also defines global evolution rules: for example gravity or temperature evolution during a year.
- Objects are used to represent some specimens that do not play a major role in the simulation. In our description, some specimens can impact the behavior of agents but their own behavior is not really interesting considering the simulation scope. So to avoid useless computation, these specimens are not described as agents but as situated objects that interact with agents.

Classic BIS provides information for environment initialization: many layers of geographic information systems are used to describe the topographic landscape and to provide objects or agents localization. Based on this layers principle, we defined the Dynamic-Oriented Modeling [11] which uses the multi-environment approach. This approach enables the splitting of the environment into sub-environments. Each one of these environments contains a specific information layer. For example, a first environment will contain topographic information, another one will be used for messages exchanges between specimens and a third one will describe the vegetal food repartition.

In GEAMAS-NG [17], our MAS platform, we also described a new temporal approach: the Temporality Model [10][16]. This model enables agents to define their own activation times and to link them with periodic behaviors. For example, an agent can define a temporality associated with its reproduction period: this temporality will periodically trigger the agent reproduction behavior. Moreover, the Temporality Model perfectly matches action's description proposed by the general part of the Descriptive Aspect of the Behavior Management Module.

Toward the building of metaverses, the main issue remains the definition of taxa's behavior. Indeed, both the Descriptive Aspect and the Representation Aspect of the BMM aims at gathering information on a particular specimen: thematicians must analyze this information to determine the taxon's generic behavior. To build a taxon's multi-agent representation, they must define its state, i.e. the key internal (physiological or intellectual) parameters that drive its behavior, its perception capacities, i.e. its sensitivity towards the environment stimuli, and its action capacities, i.e. the way it updates its state and modify its environment.

To define the complete behavior of a taxon, a first step consists in defining the set of actions this taxon can undertake. In the Behavior Management Module the Simulation Aspect is linked with the other aspect and so centered on actions description.

Defining an agent's action consists in answering three questions:

- What did trigger the action? What are the external or/and internal causes that made the agent undertake the action?
- How does the action impact the agent's internal state?
- How does the action impact the agent's environment and other agents?

Answering these questions requires knowledge of the environment the taxon evolves in, and its state. In the Multi-Agent Simulation Aspect of the BMM, we provide basic edition for action definition: having defined the key parameters of the taxon's state and the parameters of environments, the user can select some of the parameters and propose the action precondition (Figure 10). In the same way, user defines the consequences of the action on the taxon's state and the environments (Figure 15).

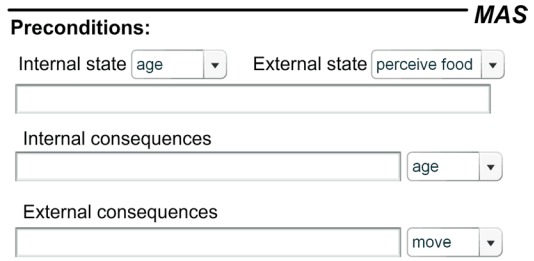

Figure 15. Description's fields of the behavior's form dedicated to graphical representations

We chose to provide a basic formalism so that thematicians can fill in the Multi-Agent Simulation aspect with a minimal assistance from modelers. The modelers, or the multi-agent system, can then adapt the behavior to more complex formalisms like DEVS [20] or Netlogo [19].

D. General considerations on the proposed solution

The three approaches used to describe the same phenomenon "action" can be identified as equivalent to the Model-View-Controller (MVC) paradigm developed in computer science. The textual description of an action corresponds to model's presentation; the graphical description's section is a view; whereas the simulation's part represents the controller. By reproducing the MVC paradigm in the frame of a BIS module, we provide a stable and adaptive structure to the immersive world's generation process.

As we have said, each of the three aspects corresponds to a specific profession: ethologist, computer graphics expert, and simulation expert. Even if each of them can limit their use of the BMM at their specific section, the real improvement in the process of porting a part of the IS as a virtual world is reached when their work can be merged in one representation: the meaning given by the ethologist, the quality of representation provided by the computer graphic expert, and the formalism by the simulation expert. To obtain the best result on a specific mirror world, a meeting between the three specialists on a co-design platform [1] is necessary to ensure the correspondence of the three aspects. Indeed, without a good communication between the ethologist and the two other experts, the consistency of the three aspects is not guaranteed. Only the expert in behavior has the scientific knowledge to accurately describe relevant movements done by specimens during specific behaviors.

However ensuring the communication between the experts is not a trivial task. In this way, multi-competent profile experts constitute a real advantage for the system, as they could check data consistency. This should be done at least for entries at the top of systematics hierarchy. Thus, the validated data can then be considered as references for users working at more specialized level of the taxonomy. In this way, it is important to feed the BMM with a set of generic initial values describing very common actions for each kingdom of systematics. These data will provide a frame for a future evolution of the module in terms of error prevention and detection. Moreover, it is important that ethologists use the same labels for equivalent actions and states at different levels of the systematics, in order to ensure the automation of the specialization and generalization process in data representation and simulation.

Alongside these recommendations that contribute to the proper deployment of the application, several improvements of this tool can be set up. More than an automatic comparison between specific and generic data, a comparison between different occurrences of equivalent actions for different taxa would greatly improve scientific knowledge in term of species' evolution. Another track of BMM technical evolution would put up with the enhancing of the interoperability between several BMM. Indeed the BMM shares its data with other modules and other BIS through webservices. Evolving the BMM to a meta-component able to exchange, compare and analyze data of other BMM would provide a controlled approach on information availability, validity, and contributor's relative participation. The more structured data is available, the better the quality of simulation and representation will be.

V. THE GENERATION OF BIS' IMMERSIVE REPRESENTATIONS

Because of the diversity and the huge amount of information contained in BIS, it is very difficult to evaluate the contribution of a user, to understand and keep in mind all the data associated to a specimen or a project. Furthermore,

this information is hardly understandable in that shape to the general public. In the same time, the progress of real time 3D technologies allows to build new representations of information by creating virtual worlds, also called "metaverses". These metaverses are built up by an aggregation of 3D models which can be seen from any point of view at any time (spatiotemporal representation). These 3D shapes can represent any object, with a variable Level Of Details (LOD). Because metaverses offer visualizations close to reality, it is more easy to figure out and understand specific configurations of the IS entities through their realistic representations [7]. Modeling research results through virtual immersive learning environment also constitutes an ideal way to analyze and communicate them with decision-makers and the general public.

However, obtaining 3D representations of information systems, and specially those dedicated to biodiversity management, is not a trivial task since the creation of the virtual world must be automated to support on demand generation. This section is a contribution to achieve this goal through a Biodi-Verse module, by defining a generic process to build metaverses from IS data [15], and establishing a typology of the different virtual world models that can be obtained.

A. From 2D to 3D BIS representation: General Steps

The principle of porting IS information from a 2D interface to a 3D immersive environment assumes that metadata associated to its entities provide a way to place (through lat-long coordinates) and to represent (3D models) them. If this last point is missing, then a substitute like generic representations (3D icons) can be used. We must precise here that our work aims to produce three-dimensional views of BIS data, not a 3D interface to manage them. The general process of the porting (Figure 16) can be divided into 3 phases:

1. The user (thematicians, decision-makers or general public) expresses his request to the system, that is to say, defines the part of the IS he wants to see as a metaverse. In order to limit the processing time, it is necessary to reduce the research's field by submitting a context of request to the user. An acceptable basis is to focus the request on a project (i.e. a metaverse representing all the entities of a particular project), or on users (i.e. a metaverse representing all entities belonging to specified users).

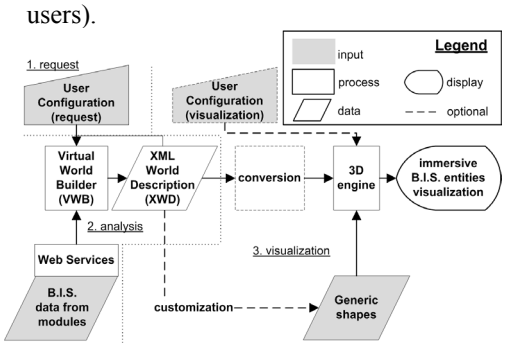

Figure 16. Data and generic processes to 3D IS

2. The description of the metaverse should be obtained by a software, the Virtual World Builder (VWB) [13], which interrogates the BIS (by using, for instance, its webservices). This analyzer works in 2 steps. First, it takes a "photo" of the IS data (depending on the above request) at a precise time by creating an XML World Description (XWD, see Illustration 1) file gathering all the instances and metadata that have to appear in the virtual world. It creates a list of the entities that matches the request's specifications. Then, the analyzer builds the metaverse architecture that will contain the entities and places them on it. The more accurate the VWB makes this generic XWD and the more accurate to the IS the metaverse will be. The configuration of the analyzer by users allows the software to determine the world's architecture that will contain the representation of entities.

```
<world name="..." builder="..." date="..." owner="...">
  <request="select * from entity where owner in ('user1', 'user2') order by owner,
  entities, collection" />
  <description>
    <room name="entrance" label="1">
        <room name="user 1 entities" label="1.1">
                ... (list of entities and metadata)...
        </room>
        <room name="user 2 entities" label="1.2">
                ... (list of entities and metadata)...
        </room>
    </room>
  </description>
</world>
```

Illustration 1. Simplified example of XWD

3. The description of the virtual world can then be combined with 3D models by the 3D engine in order to create the metaverse corresponding to the user visualization request (configuration of the VWB). Depending on the 3D engine, the XWD file is converted in a compatible input format. The specificities of the representation (level of details, representation type) are settled by the user before the metaverse generation, in order to lead and limit the complexity of the 3D world. It can be seen as a second filter, after the initial request.

As in all metaverses, a single click on a 3D entity can show the 2D card of its metadata from the IS, this representation combines the advantages of standard representation with the immersive visualization. We will now focus on the possibilities the Virtual World Builder could offer by defining the different appearances that should lead the generation process.

B. Typology of immersive representations for IS dedicated to biodiversity

Depending on the nature of entities, and the information linked to them, several representations are possible. In this part we will present the two models of representation and their declinations that correspond to four levels of details.

1) Tree of rooms

In order to represent the maximum of information from BIS, it is necessary to imagine a way to generate a metaverse compatible with all their entities, whatever the nature and the number of metadata associated to them would be. The tree of

rooms is the first model of representation, and the most general one. It consists in a schematic 3D view of several rooms linked by tunnels, forming a tree data structure. Each tunnel gives information about the following room which contains 3D icons of entities' instances.

The positioning of these rooms and the connections between them are driven by the XWD which is itself built from the user request. Because each entity has a specific nature (contact, document, taxon, specimen, observation, map, definition and project) and belong to at least one user, we assume these two points can be used as default classification parameters for the representation of the entities in virtual worlds. Most of time, instances of entities are gathered in collections and subcollections in the Information Systems' modules. This third parameter should then be suggested to the user, in order to refine the generic arrangement of the metaverse.

Thus, the user chooses in his request which restrictions could be done (for instance, build a world representing the files shared by a specific user), and the kind of hierarchy that should be used to build the XWD (example: by user, then by nature of entity, then by collection). Note that if the request aims at representing a whole biodiversity project stored in the project management module, the hierarchy is inherent to the project's structure, so it is not necessary to ask the user for classifications or hierarchy details.

Because the size of each room depends on the number of files contained in it, the general architecture of the metaverse gives a good idea of a project importance, or of users participation in the IS content.

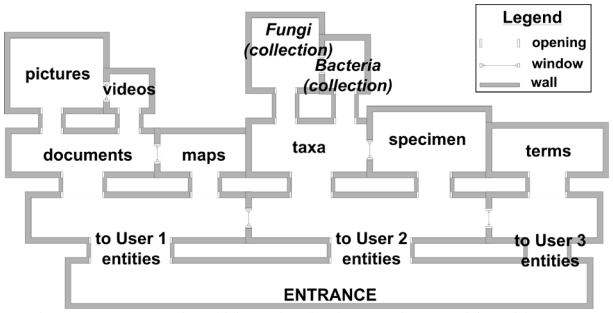

Figure 17. Example of blueprint for immersive world architecture based on the tree of rooms model

As we can see on Figure 17, there are a lot of lost spaces beside tunnels in the metaverse architecture. But that is one of the advantages of virtual architectures: usability and appropriation can be the main focus of the world's shape, independently of cost or space optimization's matters. 3D representation also helps to show several levels of granularity. Thus, if the tree data structure is used to represent groups of entities, relations between them stored in the IS can be represented by wires linking them above the room's walls (Figure 18). So users can "jump" from a document to all the related ones easily. Of course these links can be deactivated on demand in order to preserve the visibility of documents. In that way, this first level of representation offers several granularity levels corresponding to several data structures: one tree of hierarchical entities for several graphs of related documents.

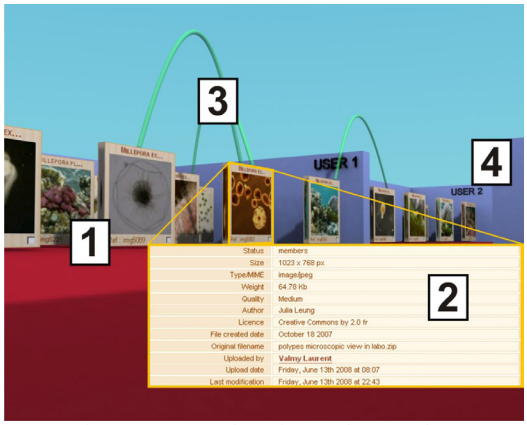

Figure 18. Screenshot of Tree of rooms view
1: 3D icons, 2: metadata, 3: links between entities, 4: path to the next room

2) Mirror Worlds

One of the specificities of IS dedicated to biodiversity is that they mainly deal with entities that have a physical materiality in the real world. In that way, it is possible to generate a virtual world trying to reproduce reality with 3D shapes correctly positioned. This kind of virtual worlds are called "mirror worlds" and constitute the second model of representation. Three increasing levels of complexity can be defined and mixed together with this model.

a) Static mirror worlds

Static mirror worlds try to reproduce reality at a precise frozen time: it is a photo of a scene reconstitution. All represented entities need a precise geolocation in order to be placed in the virtual world. Furthermore, to proceed to the automatic generation of the instantiated entities, several descriptive metadata are required to build their three-dimensional shapes. TABLE II. presents a list of entities that could be shown in mirror worlds, and the associated descriptors (other than location) that would greatly help to automate the process of shape customization.

Entity	***User***	***Specimen***	***Map***	***Sound (document)***
Common in most IS	gender	associated photo		
Not common. Improve virtual objects recognition	real dimensions (width, length, and height)		altitude of the viewpoint	scope
	skin color, haircut/color, distinctive signs (glasses/beard)	viewpoint and clipping of associated photos		
Generated Object	3D avatar	3D shape or textured pictures adapted to visualization angle	move view at a precise position and zoom	sound played at the defined position

TABLE II. ENTITIES VISUALIZATIONS IN MIRROR WORLD AND METADATA NEEDED

This level of details has two main advantages. First, because static mirror worlds tend to reproduce reality, they are directly understandable by all kind of public. Second, they can display a configuration at a precise time and place,

so it is possible to represent and immerse the user in configurations that do not exist anymore.

But mirror worlds also bring new constraints and difficulties compared to the first model:

- Mirror worlds are usually built by placing 3D models on a background map (generally an orthophotographic view of the place) as ground. So it is necessary to have a background map of the place which needs to be represented, at a scale that fits user's request.
- Depending on the request, it is difficult to make a metaverse adapted to several scales (microscale or macroscale). Without the assistance of a proper Head-Up Display (HUD), users risk to miss existing information.
- It is impossible to represent entities that have not a physical materiality (for example a definition in the thesaurus). If it should imperatively be represented in the virtual world, then it would be compulsory to use the 3D icons like those used at the tree of rooms' level.

More generally, when an entity's representation is not available at a level of detail, it is possible to use its representation at the precedent level. For instance, if a generic 3D shape is not available for a specimen, it is possible to use its photo as a planar representation.

b) Scripted mirror worlds

The scripted mirror world is the second level of this model and introduces the notion of time in the generated metaverse. Indeed, BIS contain temporal information. Especially, project management modules contain information like "specimen S of taxon T has been observed (observation O) at time tx and place P (map) by user U". Several data from a specimen studied in the frame of a monitoring experimentation constitute the script of a chronosequence that can be represented by an interpolated animation: the movement of thematicians and specimens' representation in the metaverse (Figure 19). Of course, depending the size of the gap between interpolated keys, an important bias can be introduced, but the resulting animations are still interesting for the immersion as long as this uncertainty is clearly notified to user and identified in the 3D scene (opacity modified during interpolation).

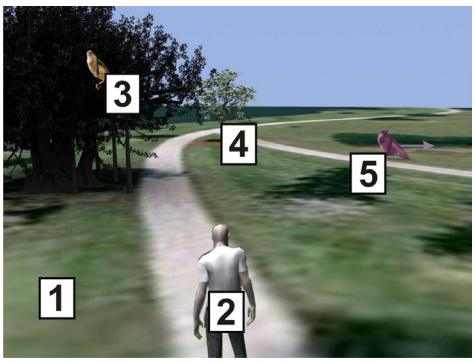

Figure 19. Screenshot of mirror world view

1: background map, 2: user, 3: planar representation of a specimen, 4: 3D representation of a specimen, 5: planar representation of a moving specimen (colored/direction given by an arrow)

With the implementation of a behavior module like the BMM in MABIS, common actions (like eating, sleeping, etc.) could then be represented. The most complex available representation is used to show the action: tridimensional shapes are better than videos, which are better than photos, themselves superior to icons in terms of immersion experience. If none are available, it is then possible to find a substitute at a more generalist level of the systematic hierarchy (Figure 20). In the case of an immersive representation, depending the choices made by the users during the visualization's configuration, a specific 2D representation (icon) can be given a better importance than a generic 3D representation.

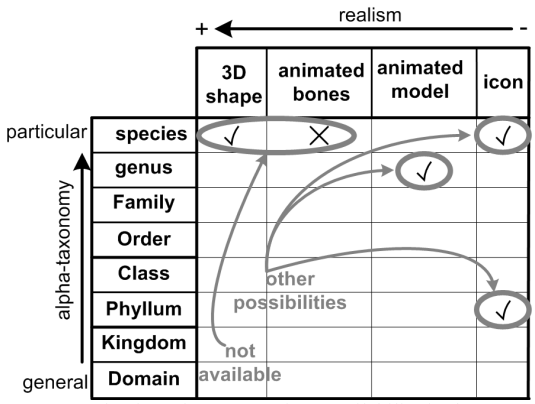

Figure 20. Evaluation of representations availability in the BMM for an action, given their precision (taxonomic hierarchy)

The animation in the virtual world can focus on a specific place or follow a particular specimen among the other inventoried in the area of study. Because this representation allows a more natural visualization of data, it is often easier to analyze than the original script. By enlarging the research scope of the VWB in the IS (from a project's entities to the whole IS), and adjusting the tuning of the passage of time, researchers can discover which other specimens and thematicians come at the same place but different times. Thus, it is possible to identify inter-species relations and establish collaborations between specialists working on nearby subjects.

So, this level of representation makes a new step toward the representation of reality, which is very useful to handle specific configurations of entities. Showing animated chronosequences is also demonstrative and useful for the analysis of collected data, but, given the focus made by IS dedicated to biodiversity, restricted to the representation of users and specimen (Directory and Systematics modules). However, it uses the information from other modules (e.g. cartographic module).

c) Simulation metaverses

The scripted mirror world has introduced the time factor by representing actions described in the project management module. The third, and highest level of complexity in this second model, keeps the time factor, but instead of just reproducing past actions, it introduces a simulation engine that analyses the scripted data of each module, and particularly those provided by the BMM, in order to

represent possible present and future of specimens' instances.

There are two ways to simulate entities' moves: by using a mathematic model, or a multiagent system [2]. In the first case, most of the actions in the simulation are established by determining periodic elements in the script. In the second case, each entity (specimen) becomes an agent in the simulation. The aim is to generate rules to define agent behavior and interactions from its data in the IS. For the moment, this can only be done by simulations' experts. Automating this step is facilitated by the structure chosen for the Behavior Management Module.

Using a multiagent approach, it is possible to link specimens to the behavior module that returns adapted rules to feed agents in the simulation. Because species behavior can be linked to their taxonomic level, generic rules can be determined to ensure results at several granularity levels. Indeed the BMM module greatly helps BIS that have to generate simulation metaverses.

C. General considerations on the proposed solution

BIS entities can be represented in metaverses as four representations corresponding to two models (Figure 21). The first representation (view) is generic, compatible with almost all IS structures, whereas other representations require a lot more metadata and specific information related to entities. However, each view has its own relevance. The first view provides an easy way to immersively evaluate the IS content according to specific grouping factors, as "user" or "project". The second view is adapted to the immersion in a frozen scene to analyze its configuration. The third view (second level of the second model) is an animated reality reproduction that tries to represent moves and actions: it is adapted to scene reconstitution. The fourth view is based on a simulation engine, and represents its output as a metaverse in order to facilitate the simulation's understanding by decision-makers.

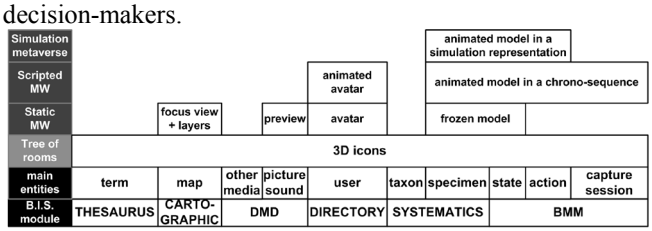

Figure 21. Immersive representations of information for the main entities, at each complexity level

If we analyze the request that allows the generation of a metaverse, we can determine that it can be divided into two steps. The first one is always to select the entities that have to be represented in the virtual world. The second step, in the first model, is to build the tree structure, whereas, in the second model, there is no structure to build but a background map to retrieve, on which entities will be placed by exploiting their geographical metadata. That fact means that, from an XWD file, it is not possible to change from a model to another without making new requests to the IS. The XWD needs to be regenerated to swap views, instant rearrangements of virtual worlds are not possible without preserving all metadata (even those not represented) associated to entities.

Thus, depending their needs and the data availability, users can choose the representation that would help them most. However, it is possible to assist them in the choice of the representation that would give them more information by using the evaluation of the metaverse's quantity of representable information. Indeed, from the initial request, the VWB is able to determine metadata on the different models of virtual worlds that could be generated: number of entities corresponding to the request, availability of metadata (georeferencing data necessary for mirror worlds, quantity of temporal information associated to entities for scripted worlds), or determination of the application's resources limitations (i.e. generic shapes for the entities that have to be represented). Using these clues, the system can suggest a less complex view if the one asked by user risks to appear empty.

Future evolutions of the application will probably greatly improve the immersion experience in BIS. Prospection fields are numerous and rely on the same technical development that edutainment software [4] have recently received, adapted to the thematic:

- Share the immersion experience on virtual online worlds [7] that could be visited by several thematicians at the same time.
- Add a pattern recognition module to the VWB in order to generate, at the same time of the metaverse, a list of interest points that should be considered by researchers, or by general public.
- Profit from the growing interoperability between BIS to provide a better integration of distributed data with data warehouses (like Google 3D warehouse [5]).

The generation of virtual worlds from IS is indeed a research field that will require numerous studies and propositions, since it is a recent field of investigation.

VI. Overall Discussion

The MABIS architecture has been instantiated in the Etic program, and more recently in its last evolution, in the Nextic project: two French environmental initiatives to build a BIS dedicated to the management of tropical information in Mascarenhas Archipelago. The screenshots we presented are extracted from these BIS which permitted us to gather feedbacks in order to improve our model up to the one we have presented here. If the global layered structure is now stable, it is however difficult to obtain feedbacks from anything else than a driven discussion with end-users. Indeed, it is not an easy task to formalize a study to globally analyze an IS that is distributed through several applications without having to evaluate each of the modules' GUI. That is why our multidisciplinary team is now associated with knowledge engineers that focus on the evaluation of the interactions between the interfaces of modules and the different types of users.

In the same time, we are also trying to improve our innovative modules, like the Biodi-Verse that tries to generate metaverses from BIS information. Now that a

reliable process has been defined and first experimentations started, we are expecting precise analysis for testing and optimizing both BMM and Biodi-Verse applications. One of the aims is to build a protocol to evaluate the relevance of metaverses, depending their structure and level of details, and their utility to the different types of users. Although we have no doubt that the complexity of virtual world generation from BIS is a brake for the non-specialists, in the frame of the Nextic project, we first prefer to focus on experts and amateurs which represent our main users at the present time.

The MABIS model offers a convenient evolution of its own structure through its modular layered architecture. In the frame of the whole BIS, major enhancement like the adding of a new entity, is supported through the development of a new module: a WCS or a WSS, depending the complexity of the new entity and its treatments. In the frame of a single module, in order to provide the three modes (main, deported, and remote) of use, functionalities are fully decoupled from the GUI. Thus enhancement through the adding of functionalities is facilitated.

However, it is important to note that each module does not require all functioning mode. In the frame of the Nextic project, we decided to separate the authentication of users and the provisioning of their data, two features initially gathered by the directory, in two modules. The idea is to keep the provisioning in the directory and extract the authentication part in a new dedicated module supporting several protocols like Lightweight Directory Access Protocol (LDAP), OpenID and Code Access Security (CAS). This evolution is to let other institutions use our BIS through their own authentication server. In this case, the MABIS authentication module does not require a main mode as this functionality will essentially be used through its webservices, in remote mode.

This generalization of webservices in MABIS also allows to expect major enhancements in terms of inter-BIS data sharing. Two cases can be considered:

1. Exchanges between two BIS build on the MABIS structure. As the systems rely on the same architecture, it is convenient to exchange and merge the information from the two platforms. Thus the search engine of all modules should soon integrate a simple checkbox allowing to extend the search of any entities to all referenced MABIS platforms.
2. Exchanges between an information system built on MABIS architecture, and another BIS. In this case, the communication between the two platforms directly depends on the existing webservices and the data models used to describe the entities. As several structures coexist (for instance Dublin Core and METS for documents, or SDD and DELTA for taxa), and because BIS concept is relatively young, it is difficult to choose and exchange among the different possibilities.

Note that the current MABIS architecture does not impose a particular data model for entities as the structure defined is more global. However, even with two MABIS systems based on different data models describing its entities, the existence of homologous webservices ensures the facilitation of exchanges.

In terms of general evolution of the MABIS structure now, we plan to prospect on two points. The first one is to try to implement parts of WSS, like the VWB of Biodi-Verse, in SaaS. Indeed new evolutions of programming languages, like Action Script 3, allow unexplored possibilities to provide on the Web traditionally installed complex software. The second point is to build an abstract presentation layer to virtually gather all the functionalities offered by the WCS and the WSS gates in a consistent, easy to interrogate, interface. These developments are part of a strategy to strengthen the exchanges between MABIS architecture and international initiatives like the Global Biodiversity Information Facility (GBIF).

VII. Conclusion and Perspectives

We have presented a new model of modular Biodiversity Information System: the MABIS architecture. These modules are components and software offering services through three different modes (main, deported and remote). MABIS provides a better flexibility on a relevant layered structure, in order to gather different types of users usually spread on several systems. Using this model, we have focused on three specialized modules:

- An evaluation component, the SIE, to associate certificates to scientific information stored in the BIS. This tool allows to follow the enhancement of data shared by specialists and amateurs through an authoritative and a community certification process. It allows users to debates on cards information with the aim to improve their content.
- A behavior management component (the BMM), to manage ethological data on specimens and taxa. The main considered entities are action, as a behavior unit, and sampling session, that represents an ordered collection of actions associated to one or more specimens. Three descriptive approaches are supported (textual, visual, and formal) to respond to the needs of three different experts (ethologists, computer graphics experts, simulation experts).
- A Biodi-Verse software, to generate immersive representations of selected MABIS entities. Because information systems dedicated to biodiversity gather sets of data difficult to appropriate, we present several possibilities to represent them through virtual worlds. Using a generic process, it is possible to generate up to four different representations of an IS part, depending on the available resources and the data structure.

The instantiations of MABIS architecture through the Etic program and Nextic project have already given positive feedbacks that also suggest enhancement tracks to the model in terms of data exchanges between BIS. Our future researches will be directed toward these fields.

ACKNOWLEDGMENT

The ETIC program promotes the enhancement of insular tropical environment contents from biological researchers using Information and Communication Technologies. The European Union, the French National Government, and the Regional Council of Reunion funded it in the frame of the DOCUP-FEDER 2002-2006, A9-04 ICT measure.

NEXTIC is the next ETIC project for distributing Biodiversity edited contents on the Web with webservices, then with Semantic and Immersive services. It is supported by the PO-FEDER 2007-2013 measure.

We are also grateful to Mr Laurent Cochet, engineer on Nextic project, for the implementation information provided to our research team.

REFERENCES

[1] N. Conruyt, P. Conruyt, O. Sebastien, "A methodology for Designing E-services from a Co-Design Platform," Proc. of the 6th International Conference on Knowledge Management, I-know'06, Graz, Austria, K. Tochtermann / H. Maurer Eds., pp. 226-230, September 2006.

[2] N. Conruyt, D. Sebastien, R. Courdier, D. David, N. Sebastien, T. Ralambondrainy, "Designing an Information System for the preservation of Insular Tropical Environment in Reunion Island. Integration of Databases, Knowledge Bases and Multi-Agent Systems by using Web Services", Advanced Agent-Based Environmental Management Systems, Whitestein Series in Software Agent Technologies and Autonomic Computing, Cortés, Ulises; Poch, Manel (Eds.), Birkhäuser, pp. 61-90, March 2009.

[3] J. Ferber, "Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence", Addison-Wesley Professional, 1999.

[4] F. Garzotto, M. Forfori, "Hyperstories and social interaction in 2D and 3D edutainment spaces for children", Proceedings of the seventeenth conference on Hypertext and hypermedia table of contents, Odense, Denmark, pp. 57-68, 2006.

[5] A. Hudson-Smith, "Digital Urban – The Visual City", UCL Center for advanced spatial analysis, Working paper series, paper 124, 1999, ISSN 1467-1298.

[6] K. D. Fook, A. M. V. Monteiro, G. Câmara, "Web Service for Cooperation in Biodiversity Modeling", Advances in Geoinformatics, Clodoveu A. Davis Jr., Antônio Miguel V. Monteiro (Eds.), Springer Berlin Heidelberg, pp. 203-216, November 2007.

[7] G. McArdle, "Exploring the Use of 3D Collaborative Interfaces for E-Learning," Studies in Computational Intelligence, Intelligent Systems and Technologies, H.-N. Teodorescu et al. (Eds), Springer Berlin Heidelberg, pp. 249-270, July 2009.

[8] I. Marmaridis, S. Griffith, "Metaverse Services: Extensible Learning with Mediated Teleporting into 3D Environments," Information Processing, Information Systems: Modeling, Development, and Integration, Proc. International United Information Systems Conference, UNISCON, J. Yang et al. (Eds.), Springer Berlin Heidelberg, pp. 229-239, April 2009.

[9] S. J. Mayo et al., "Alpha e-taxonomy: responses from the systematics community to the biodiversity crisis," Kew Bulletin vol. 63: 1–16, Springer Netherlands, pp. 1-16, May 2008, doi: 10.1007/s12225-008-9014-1.

[10] D. Payet, R. Courdier, T. Ralambondrainy, N. Sebastien, "Le modèle à Temporalité : pour un équilibre entre adéquation et optimisation du temps dans les simulations agents, " Proc. Journées Francophones des Systèmes Multi-Agent, 2006.

[11] D. Payet, R. Courdier, N. Sebastien, T. Ralambondrainy, "Environment as support for simplification, reuse and integration of processes in spatial MAS," Proc. Information Reuse and Integration, pp. 127-131, September 2006.

[12] K. R. Popper, "All life is problem solving," Routledge (Eds.), 1999.

[13] D. Sebastien, "Contribution à la génération de représentations immersives des données de Systèmes d'Information sur la Biodiversité, " PhD. Thesis, University of Reunion Island, (to appear), 2010.

[14] D. Sebastien, N. Conruyt, "Online multimedia database for communities of practice in Biology: a real use challenge," Proc. International Conference on Internet and Web Applications and Services, pp. 549-554, IEEE Computer Society, May 2008, doi: 10.1109/ICIW.2009.89.

[15] D. Sebastien, N. Conruyt, R. Courdier, T. Tanzi "Generating Virtual Worlds from Biodiversity Information System: requirements, general process and typology of the metaverse's models," Proc. of the Fourth International Conference on Internet and Web Applications and Services, Venice, Italy, May 2009.

[16] N. Sebastien, R. Courdier, D. Hoareau, M. P. Huget, "Analysis of temporal dependencies of perceptions and influences for the distributed execution of agent-oriented simulations," Proc. European Simulation and Modelling Conference, October 2008.

[17] N. Sebastien, "Distribution et Parallélisation de simulations Orientées Agent, " PhD. Thesis, University of Reunion Island, 2009.

[18] F. Recknagel, "Ecological informatics: Current scope and future directions," Proc. Information Technologies in Environmental Engineering, Springer Berlin Heidelberg, pp. 3-22, May 2009, doi:10.1007/978-3-540-88351-7.

[19] S. Tisue, U. Wilensky, "NetLogo: A Simple Environment for Modeling Complexity," Proc. International Conference on Complex Systems, 2004.

[20] B. P. Zeigler, "DEVS representation of dynamical systems: event-based intelligent control," Proc. of the IEEE, vol. 77, n°1, pp72–80, January 1989.

ASSOLO: an Efficient Tool for Active End-to-end Available Bandwidth Estimation

Emanuele Goldoni
University of Pavia
Department of Electronics
27100 - Pavia, Italy
emanuele.goldoni@unipv.it

Giuseppe Rossi, Alberto Torelli
University of Pavia
Department of Computer Eng. and System Science
27100 - Pavia, Italy
giuseppe.rossi@unipv.it, alberto.torelli01@ateneopv.it

***Abstract*—End-to-end available bandwidth estimation is a crucial metric for bandwidth-dependent services such as multimedia streaming, peer-to-peer and gaming applications; it is also useful for quality of service verification and traffic engineering. This paper presents the details of ASSOLO, an efficient active probing tool for estimating the available bandwidth of a network path. The tool is based on the well-known concept of "self-induced congestion", and it features a new probing traffic profile called REACH (Reflected ExponentiAl Chirp) to test a wide range of possible rates with a single stream of packets. In addition, the program runs inside a real-time operating system and uses some de-noising techniques to improve the measurement process. Experimental results show that ASSOLO outperforms pathChirp, a state-of-the-art measurement tool, estimating the available bandwidth with greater accuracy and stability in presence of different cross-traffic sources. Moreover, we demonstrate that the use of a real-time operating system can increase the stability of the estimations lowering the impact of software context switches.**

***Keywords*-Available bandwidth, active network measurement, performance evaluation, real-time.**

I. Introduction

ASSOLO is a novel tool for available bandwidth estimation in packet-switched networks which has been originally introduced in [1]. This work extends some of the results presented in the original paper by investigating the performance of our tool in presence of poissonian cross-traffics. We also study the actual impact of a real-time operating system on the measurement process, and we provide more details on the filtering technique implemented into the program.

The *available bandwidth* of a network path is a crucial metric in quality-of-service management, traffic engineering or congestion control. Voice over IP (VoIP), peer-to-peer and video-streaming are examples of widely-used applications that could greatly benefit from the knowledge of the available bandwidth along an Internet path. For example, in [2] and [3] the importance of the available bandwidth is investigated respectively for peer-to-peer (P2P) networks and gaming-on-demand services. In [4] the authors focus instead on improving the perceived quality of video streaming through a dynamic path selection based on the measurement of network-layer metrics. Similarly, in [5] the authors propose a live broadcast platform where the video source is distributed to a number of clients organized in a peer-to-peer tree-structured overlay network. In this network the root node is also responsible for organizing and maintaining the position of each peer within the tree according to the available bandwidth and the latency between peers. The knowledge of the actual available bandwidth is also exploited in [6] to improve video streaming rate- and quality-adaptation decisions; results obtained through simulations show that an estimation algorithm can substantially increase streaming performance.

The same approach is also adopted in existing commercial products: Microsoft Windows Media Server includes a technology called Intelligent Streaming for on-demand and live media streaming over IP. This solution identifies the actual maximum throughput allowed by the network path using a end-to-end client/server system. This value is used to choose the best encoding rate which maximizes the quality of received media without overloading the network [7].

In principle, it would be possible to obtain estimates of the available bandwidth directly from intermediate routers along the network path; however, this is not feasible in practice due to technical and security reasons. Therefore, researchers have proposed several end-to-end measurement algorithms which infer the network characteristic transmitting a few packets and observing the effects of intermediate routers or links on these probe frames. Examples of probing tools which have emerged in recent years are IGI [8], Spruce [9], Pathload [10], TOPP [11], [12], pathChirp [13], FEAT [14] and BART [15]. They differ mainly in the structure of probe streams and in the algorithms used to estimate available bandwidth from the received packets. Nevertheless, producing reliable estimations in real-time still remains challenging: the measurement process should be efficient, accurate, non-intrusive and robust at the same time. Moreover, the algorithm should adaptively apply to different types of networks and cross-traffics, and must be able to produce fast periodic estimations in order to track bandwidth fluctuations. As a result, as noted in [16]–[18], current available bandwidth estimation techniques and tools are far from being ready to be applied in many applications and scenarios.

Compared to the tools mention above, our novel tool ASSOLO (Available-bandwidth Smart Sampling On-Line Tool) features a new probing traffic profile called REACH (Reflected ExponentiAl Chirp). A REACH tests a wide range of rates

and is more accurate in the center of the probed interval. Moreover, ASSOLO uses a combination of new and existing filtering techniques to improve the accuracy and stability of results. Finally, our tool runs inside a real-time operating system in order to minimize the impact of context switches on the measurement process.

The rest of the paper is organized as follows. In Section II we introduce the related work on available bandwidth estimation and we focus on the Probe Gap Model, the general measurement scheme adopted by our tool. Next, Section III illustrates the algorithm used to generate the REACH probing stream and the additional features introduced in our tool. An evaluation of ASSOLO is presented in Section IV, including results obtained comparing our solution to the state-of-the-art tool pathChirp both in terms of intrusiveness and accuracy. Finally, in Section V we conclude and we outline future works.

II. Related Work

Techniques for end-to-end available bandwidth estimation can be divided into two categories: active probing and passive measurements. The latter infer the required information from existing data transmissions while active probing techniques produce an estimation injecting dedicated probe traffic into the network.

Passive measurements do not require dedicated packets to perform the estimation: useful information is obtained from traffic originated by active connections providing a particular service. In this context, the idea of using TCP for network measurements has attracted a lot of studies: RTT values [19] or ACK arrival times [20], [21] have been used on the sender's side to infer the available bandwidth from existing transmissions. These methods are lightweight and fast but they can be applied only to network paths that have recently carried traffic. Moreover, congestion control algorithms, buffers and competing connections may influence the achievable throughput of a single TCP connection, thus altering the accuracy of estimations [22].

Active measurement techniques use probe packets to measure the end-to-end delays introduced by existing cross-traffic (Figure 1). These methods require instrumentation at both ends of the path; moreover, the probe traffic injected into the network may affect the performance of other applications and actually alter the available bandwidth. In addition, some tools require a long measurement time and use hundred of packets before producing an estimation. The majority of existing tools belong to the Probe Gap Model (PGM) or the Probe Rate Model (PRM).

In the Probe Gap Model, a tool sends a single probing pair or train; it exploits then the dispersion of packets on the receiver side to calculate the available bandwidth. The main assumption of this model is that the link with the minimum available bandwidth is also the link having the minimum capacity. This is probably the biggest limit of this approach: the hypothesis is not valid for many Internet paths and can results in significant underestimations of the available

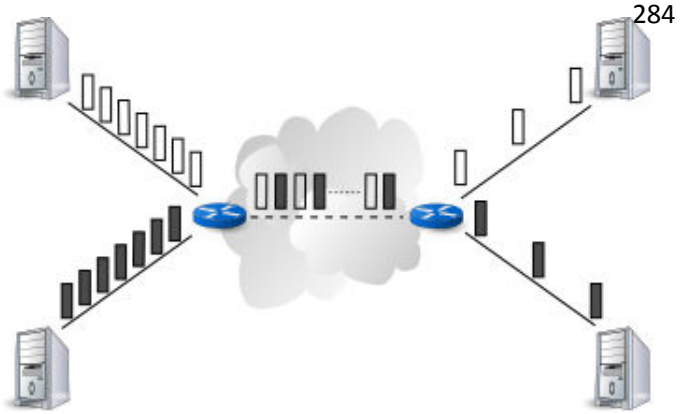

Fig. 1. The spacing effect on multiple traffics over a congested network path.

bandwidth over multi-hop links [23]. Notable tools based on the Probe Gap Model are Spruce [9], IGI [8] and Delphi [24].

Delphi [24] assumes a multi-fractal model for the cross-traffic. The main idea in this tool is that the spacing of two probing packets at the receiver can provide an estimate of the amount of traffic at a link. Spruce [9] is based too on direct probing and it uses tens of packet pairs to collect available bandwidth estimations. The input rate of pairs is chosen to be roughly around to the capacity of the path, which is assumed to be known. Moreover, packets are spaced with exponential intervals to emulate a poissonian sampling process. IGI [8] uses a sequence of about 60 unevenly space packets to probe the network and the gap between two consecutive packets is increased until the average output and initial gaps match.

The Probe Rate Model, instead, is based on the concept of self-induced congestion. The underlying idea is quite simple: if a sequence of packets is sent at a rate lower than the available bandwidth along the network path, then the arrival rate of packets at the receiver will not exhibit any notable variation and it will match with the sender's rate. On the other hand, if the sending rate exceeds available bandwidth, one or more intermediate queues will fill up and the probe traffic will experience delays. Thus, the measurement is performed through the research of the turning point at which the probe stream starts seeing an increasing trend. The PRM model has proved to be accurate and it is used in many estimation tools, such as TOPP, Pathload, pathChirp, FEAT and BART.

TOPP [11], [12] and Pathload [10] use a constant bit-rate stream, sending pairs or trains of packets at a given rate and changing this rate every round. TOPP increases linearly the sending rate in successive streams, trying to find out the exact turning point. Pathload on the other hand varies the probing rate using a binary search scheme and the final output, result of multiple measurements, is a variation range rather than a single estimate. Since multiple trains are required to produce a single estimation, the intrusiveness of these techniques is quite high and the measurement process is time-consuming.

PTR [8] is an active probing algorithm which sends several probing packets to detect background traffic. The method com-

pares the time interval at the source with that of destination and then uses the timings to estimate the value of available bandwidth.

pathChirp [13] sends a variable bit-rate stream called *chirp*, which consists of exponentially spaced packets. A chirp allows to probe the network path over a wide range of rates injecting only one stream – if the delays show an increasing trend starting from a particular packet, the associated rate is used to infer the unused capacity. pathChirp can estimate available bandwidth sending only one chirp: this feature makes the measurement process fast and lightweight. However, pathChirp samples the lower rates more frequently than the higher rates. Therefore the tool is less accurate if actual available bandwidth is not located nearby the beginning of the probing range. Smoothed-chirp (S-chirp) is a similar approach based on iterative probing and originally proposed in [25].

BART [15] relies on sequences of packet pairs sent at randomized rates. This tool uses also a Kalman filter to track the evolution of available bandwidth in real-time and to filter out noisy observations. BART is lightweight, efficient and non-intrusive; however, the tool is still in development and it is not freely available. MR-BART is a extension of the original BART method which employs multi-rate probe packet sequences to achieve faster convergence and more accurate estimations.

FEAT [14] is a recent tool which features a probe pattern called *fisheye stream*. A fisheye stream consists of packets of equal size which are sent at a changing rate, from a lower bound to a maximum probing rate. The tool identifies also an interval, called "focus region", where the available-bandwidth is most likely to be. Inside this region the sampling frequency is higher and the number of packets sent for each sampling rate is larger. This approach creates a more identifiable turning point but it also makes the measurement process intrusive.

While BART and FEAT look quite promising, it is difficult to compare them to other state-of-the-art tools: the results presented by the respective authors have been obtained only through simulations or using specific Internet paths, and to our best knowledge the two programs have never been released publicly.

III. ASSOLO

ASSOLO is an available bandwidth estimation tool which has been originally presented in [1]. Unique to this tool is a new probe traffic profile called REACH (Reflected ExponentiAl Chirp), which tests a wide range of rates using a single stream of packets and injecting a negligible amount of traffic into the network. The tool introduces also some techniques to minimize the impact of different sources of errors on the estimation process.

A. Probing stream

ASSOLO is based on the concept of "self-induced congestion" – it tests different rates using a single stream of packets, and then infers the available bandwidth harnessing the information about the relative delays. This approach has a twofold advantage: it requires neither clock synchronization nor clock-offset knowledge between the two end-hosts probing the network. However, it is important to consider that the first packet of the train itself does not have any associated rate. Instead, it is used as a reference value to calculate all successive *relative* queuing delays within a stream.

The novel REACH probing traffic profile tests multiple rates with a single stream, and it is more accurate at the center of the stream, where the actual available bandwidth is likely to be. A similar idea was originally proposed in [14], but our method introduces a different spacing algorithm and sends less packets. Compared to pathChirp, the stream used by ASSOLO is different too – both tools use a sequence of packets with increasing delays, but the shape of the traffic and the delays within a stream are not the same.

The REACH stream used by our tool tests different rates increasing the instantaneous packet rates from a lower bound L to a maximum rate U. The first k packets of the stream probe values lower than the center $H = \frac{U+L}{2}$; additional k packets test values between H and the maximum probing rate U. However, the probing rates do not increase linearly in a REACH. Instead, the density of the stream increases as well as values approach the center of the interval $[L, U]$. Then, once the rate H has been tested, the probing density start decreasing. The same can be said for the accuracy of the estimation, since it is proportional to the density of the probing stream.

The maximum relative accuracy of ASSOLO's estimations is defined by the parameter σ. Given the probing range, the absolute error S around the center of the probing interval is calculated as:

$$S = \sigma \left(\frac{U - L}{2} \right). \tag{1}$$

Moreover, the algorithm uses a coefficient γ to control how fast the density of streams changes. This parameter reminds the *spread factor* used by pathChirp, although the two resulting trains are quite different. ASSOLO uses by default $\sigma = 5\%$ and it sets γ to 1.2. However, it is important to note that the choice of these parameters is arbitrary – values should be assigned according to the specific requirements of the target application. Decreasing γ and σ, the tool would send more packets but it should result in a more accurate estimation; similarly, increasing these value should reduce both intrusiveness and accuracy.

An additional parameter Δ_x is also needed to better describe the REACH stream generated by ASSOLO. The function of this auxiliary coefficient is to describe the gap between two consecutive packets of the stream, and it is defined as follows:

$$\Delta_x = S \cdot \gamma^{|x-1|} \tag{2}$$

and combining Equations 1 and 2 we get:

$$\Delta_x = \sigma \frac{U - L}{2} \gamma^{|x-1|} \tag{3}$$

Starting from the center H of the probing interval towards the upper bound U, instantaneous packet rates in a REACH are

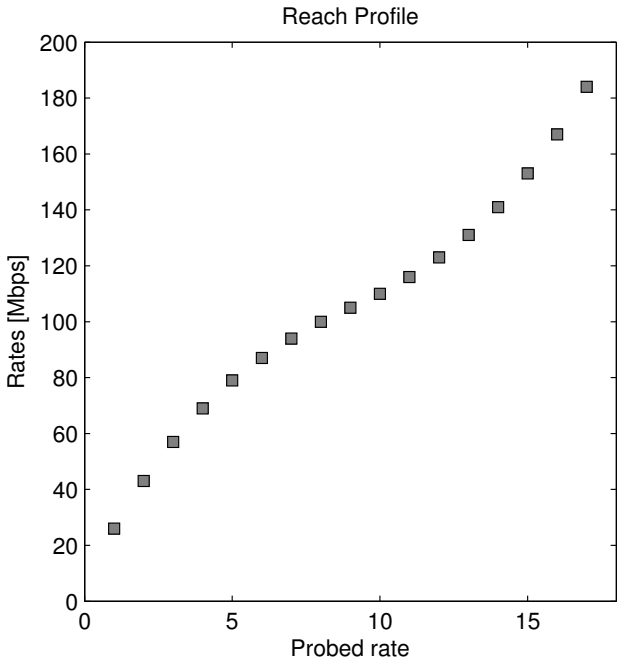

Fig. 2. Distribution of packets in a REACH stream.

$H, H+\Delta_1, H+\Delta_1+\Delta_2, H+\Delta_1+\Delta_2+\Delta_3, ...$ A more formal description of instantaneous probing rates R_x tested by this stream is:

$$\begin{cases} R_x = H, & \text{if } x = 1 \\ R_x = R_{x-1} + \Delta_x, & \forall x > 1, R_x < U \end{cases} \quad (4)$$

On the other hand, probing rates from the center towards the lower bound are $H, H-\Delta_1, H-\Delta_1-\Delta_2, H-\Delta_1-\Delta_2-\Delta_3, ...$ The instantaneous rates tested by a REACH can then be described as:

$$\begin{cases} R_y = H, & \text{if } y = 1 \\ R_y = R_{y-1} - \Delta_y, & \forall y > 1, R_y > L \end{cases} \quad (5)$$

The resulting stream is shown in Figure 2. As also the name REACH (Reflected ExponentiAl CHirp) suggests, the profile is symmetric: the right and the left part look like two mirrored exponential functions.

Since the function is symmetric, we can analyze only the right part of the stream – the same considerations would also apply to the left one. ASSOLO uses k packets to test values between H and the upper bound U. Thus, the instantaneous rate R_k associated with the k^{th} packet is the maximum probing rate. We can write this condition as:

$$U = R_k = H + \sum_{i=1}^{k} \Delta_i \rightarrow U - H = \sum_{i=1}^{k} \Delta_i \quad (6)$$

If we substitute the values of Δ_i and H, we get:

$$U - H = \sigma \frac{U-L}{2} \sum_{i=1}^{k} \gamma^{|i-1|} \quad (7)$$

$$U - \frac{U+L}{2} = \sigma \frac{U-L}{2} \sum_{i=1}^{k} \gamma^{|i-1|} \quad (8)$$

$$\frac{U - \frac{U+L}{2}}{\sigma \frac{U-L}{2}} = \sum_{i=1}^{k} \gamma^{|i-1|} \quad (9)$$

$$\frac{\frac{U-L}{2}}{\sigma \frac{U-L}{2}} = \sum_{i=1}^{k} \gamma^{|i-1|} \quad (10)$$

$$\frac{1}{\sigma} = \sum_{i=1}^{k} \gamma^{|i-1|} \quad (11)$$

In addition, the value of the truncated sum is:

$$\sum_{i=1}^{k} \gamma^{|i-1|} = \frac{\gamma^{k+1} - 1}{\gamma - 1} \quad (12)$$

Combining Equations 11 and 12, we get:

$$\gamma^{k+1} = \frac{\gamma - 1}{\sigma} + 1 \quad (13)$$

which leads to

$$k = log_\gamma \left(\frac{\gamma - 1}{\sigma} + 1 \right) - 1 \quad (14)$$

Actually we should define R_k as the maximum sending rate *not exceeding* the upper bound U. Equation 6 should then take the form:

$$U \geq H + \sum_{i=1}^{k} \Delta_i \quad (15)$$

Hence the correct value of k is:

$$k = \left\lfloor log_\gamma \left(\frac{\gamma - 1}{\sigma} + 1 \right) - 1 \right\rfloor \quad (16)$$

As we mentioned before, a REACH uses the first k packets of the stream to probe values lower than the center $H = \frac{U+L}{2}$. Then, the stream probes the rate H; finally, other k packets tests values between H and the maximum probing rate U. As a result, a REACH probes $2k+1$ rates exploiting relative delays between probe packets. Therefore, our tool needs to send an additional packet at the beginning of the REACH. The total number N of packets used by ASSOLO to probe $2k+1$ rates is:

$$N = 1 + (2k+1) = 2k+2 \quad (17)$$

Since we know the size of a REACH, we can also combine Equations 4 and 5 and describe the rates probed by a REACH profile as:

$$R_j = H + \text{sign}\left(j - \frac{N}{2}\right) \cdot S \cdot \frac{\gamma^{|j - \frac{N}{2}|} - 1}{\gamma - 1} \quad (18)$$

B. End-hosts predictability

A fundamental difficulty with the existing measurement tools stems from a number of issues on both end-hosts and network paths [26]: system timing, hardware errors and end-to-end pathologies could produce a considerable amount of noise in the individual network observations. For example, the Linux kernel is a time sharing operating system designed to give a fair share of the CPU in a multi-user environment [27] – even some kernel services like memory allocation and system calls exhibit some non-deterministic timing behavior.

Network measurement tools have strict operational deadlines between the arrive of a packet and the application's response to that event – the same can be said for the sender side, where the packet sent by the application should ideally start with no delay. In [28] the impact of context switching on the measurement process is analyzed in depth. Some tests conducted in our lab confirmed that a significant amount of noisy observations is due to the non-deterministic behavior of the operating systems hosting the sender and the receiver. To accommodate deadlines on both the end-hosts we decided to use a real-time operating systems (RTOS), which can guarantee predictability and accurate system timings for applications. ASSOLO runs inside a GNU/Linux system with RT-Preempt [29], [30] patch enabled, thus using a fully preemptible kernel with high-resolution kernel timers. In order to minimize the impact of context switches on the bandwidth estimation process, the tool gets the highest priority on both the end-host systems while probing the network.

Our program could be easily ported to other real-time operating systems, since it is written in C language and uses standard system calls. However we decided to use the RT-preempt approach, which makes the software much more portable and easier to deploy and maintain over a large network infrastructure.

Compared to other Linux real-time approaches, such as RTAI [31] and Xenomai [32], RT-Preempt is not a hard real-time approach in strict sense: processes can incur a latency that is not deterministic and no guarantees are usually provided on the feasibility of a given task set. Although this real-time extension to the Linux kernel suffers from the above-mentioned limitations, it greatly improves the performances of many applications and the responsiveness of the whole system, thus providing adequate service for most applications that need real-time determinism [33]. Moreover, no special programming libraries are required: the applications compiled for RT-Preempt Linux can be also used on a standard, non real-time Linux system with negligible adaptations.

C. Observations filtering

Like the end-hosts, also intermediate routers can be heavily affected by predictability issues: interrupt coalescence, clock resolution and context-switching delays are all factors that can potentially modify timings of the probe traffic, therefore introducing errors. Moreover, almost all existing tools assume the hypothesis of fluid cross-traffic [34], ignoring the discrete nature of packets. However this non-deterministic behavior of intermediate nodes depends on the specific network path and it can not be easily controlled or even described. As a result, most of existing available bandwidth estimation techniques produce noisy observations [35], [36].

A vast majority of available bandwidth estimation tools introduce filtering techniques: for example, Moving Average, Exponential Weighted Moving Average (EWMA), Wavelets or Kalman filters have been successfully adopted in [13], [15], [37]–[40] to attenuate noise and local random fluctuations, converting noisy values into a reliable estimate.

The idea of using such a solution in this context is based on the predictability and long-term stability of the Internet. Typically, the available bandwidth of an Internet path shows strong correlation and a certain degree of stability over intervals that span from several minutes to a few hours [14], [41]. Given a new observation, an effective filtering technique can produce a new estimate of the available bandwidth combining both the most recent observation and the old values.

For example, the Exponentially Weighted Moving Average (EWMA) filter uses one or more observed values O_k and outputs a new estimation E_i calculated as follows:

$$E_i = \alpha E_{i-1} + (1-\alpha)O_i. \tag{19}$$

This filter is used by some estimation tool like Abing [37] and Yaz [38]. However, the difficulty with the EWMA technique lies in the choice of the exponential weight α. With large values of α, the old estimates are given more importance and the filter is slow but stable; agility is instead achieved by keeping α small. Ideally, the filter should be adaptive, setting the value of α according to the current circumstances: sharp and non-persistent changes can at first be treated as noise using lower weights α_i. However, if the change persists, the filter should quickly converge to the new value. Equation (19) should then take the form:

$$E_i = \alpha_i E_{i-1} + (1-\alpha_i)O_i. \tag{20}$$

Lowpass EMA [42], *Stability* [43] and *Error Based* Filters [43] are three existing techniques designed around this philosophy. Although they have been proposed a couple of years ago, to our best knowledge none of them has actively been employed in an available bandwidth estimation tool.

In [44] we originally proposed the use of Vertical Horizontal Filter (VHF) in such a context. The VHF filter is a modified EWMA technique borrowed from the financial world [45] which can dynamically modify its behavior according to trends identified in the temporal evolution of available bandwidth according to the same principles of the three above mentioned filters. The dynamically exponential weight α_i in (20) is computed as:

$$\alpha_i = \beta \frac{\Delta_{max}}{\sum_{t=i-M}^{i} |O_t - O_{t-1}|} \tag{21}$$

where Δ_{max} is the gap between the maximum and the minimum values in the M most recent observations. We set β as $\frac{1}{3}$ and the window size $M = 10$, although these parameters

were obtained empirically and a careful choice could bring further improvements.

We performed a series of simulations to investigate the effectiveness of different filtering techniques on the available bandwidth estimation process. Compared to the methods mentioned above, we found that the VHF filter leads to better results in many cases and shows greater stability. Our experiments also indicated that there is no need to fine tune the VHF filter every time some network conditions change. A detailed description of VHF and a comparison between different linear filtering techniques can be found in [44].

Results persuaded us to employ the Vertical Horizontal Filter, which is used inside our tool to cope with noisy observations and to estimate the actual available bandwidth from raw measurements.

D. Excursions segmentation

According to the basic principle of PRM's self-induced congestion, an instantaneous sending rate higher than the actual available bandwidth results in increasing queuing delays at receiver; otherwise, packets sent by a tool will experience no delays. This model is valid also for tools which probe multiple rates with a single train, like ASSOLO does – the last instantaneous probing rate which does not result in an increasing queuing delay is considered a simple estimate of available bandwidth. However, this approach oversimplifies reality, lacking, for example, to consider cross-traffic bursty behavior and end-host interrupt coalescence effects.

Traditional network adapters generate an interrupt for each received frame, thus generating up to thousands of internal signals per second in high-speed networks. These interrupts consume a lot of system's resources and introduce a significant amount of context switches, resulting in a CPU overhead. [46]

To mitigate the effects of this issue, some network adapters recently introduced the support for Interrupt Coalescence (IC) [47]. This solution decreases the processing overhead buffering multiple packets before generating a single interrupt for the burst of frames. A similar approach has been introduced in NAPI [48], a modification to the device driver packet processing framework of Linux kernel. NAPI mixes interrupts with a polling approach to implement an adaptive interrupt coalescing which modifies its behavior according to the actual network load. This solution usually results in improved performances for high-speed networking. Although IC decreases the per-packet processing overhead, it introduces also non-deterministic queueing delays, thus altering the time spacing of packets in a probing train. As noted in [49], IC can be detrimental to TCP self-clocking making the traffic more bursty, and it has a negative effect on the accuracy of active and passive bandwidth measurements.

The typical profile of queuing delays in a train is often non-monotonic. For example, Figure 3 shows the typical queuing delays of a chirp sent by pathChirp: one or more excursions produced by bursts return to zero, while a final excursion ends with increasing queuing delays.

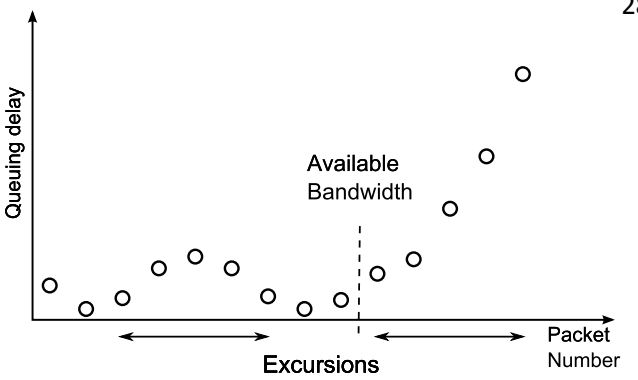

Fig. 3. Typical queuing delays in a chirp.

function EXCURSION(q, i, F, L)
  $j \leftarrow i + 1$
  $q_{max} \leftarrow 0$
  while $(j \leq N)\, AND\, (q_j - q_i > q_max/F)$ **do**
    j++ ▷ Count excursion's packets
  end while
  if $j \geq N$ **then**
    return j ▷ Non-ending Excursion
  end if
  if $j - 1 \geq L$ **then**
    return j ▷ Excursion
  else
    return i ▷ Not an excursion
  end if
end function

Fig. 4. Pseudo-code for the pathChirp's excursion segmentation algorithm [13].

The authors of pathChirp introduced a smart segmentation algorithm to cope with this kind of burstiness effects, detecting increasing delays belonging to a cross-traffic bursty transient. The main goal of pathChirp's excursion segmentation algorithm is to identify potential starting and ending packet i and j respectively for an excursion. Potentially, every packet i where queueing delay q_i starts increasing could be a starting point of an excursion. We define the end of the excursion as the point where the queuing delay returns to zero or where it has decreased by a factor F from the maximum queueing delay experienced during this interval. Moreover, if the distance between these two packet is long enough, for example longer than a threshold L, then all packets between i and j form an excursion. On the other hand, the last excursion identifies the congested region and it does not terminate. The pseudo-code of the procedure is presented in Figure 4 while a detailed description of the whole algorithm can be found in the original paper of pathChirp [13]. Since this solution proved to be quite effective to cope with burstiness, ASSOLO adopts exactly the same technique to analyze the queuing delays of each single REACH and to identify the correct turning point.

E. Availability

Additional implementation details and a copy of the source code of ASSOLO are all freely available at *http://netlab-mn.unipv.it/assolo/* or through the authors. Future developments and data reports will be published at the same location.

IV. Results

In order to evaluate our estimation method, the performance of ASSOLO has been studied in a controlled testbed environment. In addition, we compared the intrusiveness and the accuracy of our solution with pathChirp, a similar state-of-the-art measurement tool, in presence of poissonian or constant bit rate (CBR) cross-traffics.

The testbed configuration is shown in Figure 5. Two computers using Ubuntu GNU/Linux are connected together through a Fast Ethernet cross-cable and serve as routers. Two other machines of the testbed simulate a source of controlled traffic flows using the D-ITG tool [50], which loads the network generating synthetic flows of known properties and statistical distributions. Finally, the sender and the receiver for each measurement tool use additional PCs running Ubuntu GNU/Linux with a standard or real-time kernel. Prasad et al. in [51] showed that each store-and-forward device introduces an additional serialization latency in a packet's delay. This can result in a consistent underestimation of the hop's capacity. Therefore, we provisioned the network with two Fast Ethernet switches in order to introduce an additional potential source of errors during tests.

The topology of the testbed is quite simple but sufficient to evaluate the performance of a measurement tool: for example, the same configuration has been used in [52] and [17] to perform an experimental comparison of different available bandwidth estimation tools.

We adopted the default configurations for both probing tools: ASSOLO uses $\sigma = 5\%$ and γ to 1.2 while the γ of pathChirp has been initially set to 1.2. Since results obtained in [13] showed that pathChirp generally performs better with larger packets, the packet size for both tools was 1000 byte. Finally, the upper and the lower bandwidth bounds U and L were respectively equal to 200 and 10 Mbps; however, both tools automatically adjust the values if the range is too narrow. A complete list of all the configuration parameters of testbed's devices and tools is provided in [53].

A. Intrusiveness

The intrusiveness of pathChirp and ASSOLO can be easily compared. From [36] we know that a *chirp* is composed of N packets, where N can be calculated as follows:

$$N_{chirp} = \left\lfloor 2 + \frac{1}{log\gamma} log\left(\frac{U}{L}\right) \right\rfloor .$$

The size of the stream sent by pathChirp depends on the upper (U) and lower (L) rate bounds. However, pathChirp automatically reduces or increases the probing range if it is too wide or too narrow: as a result, the tool sends on average 15-20 packets.

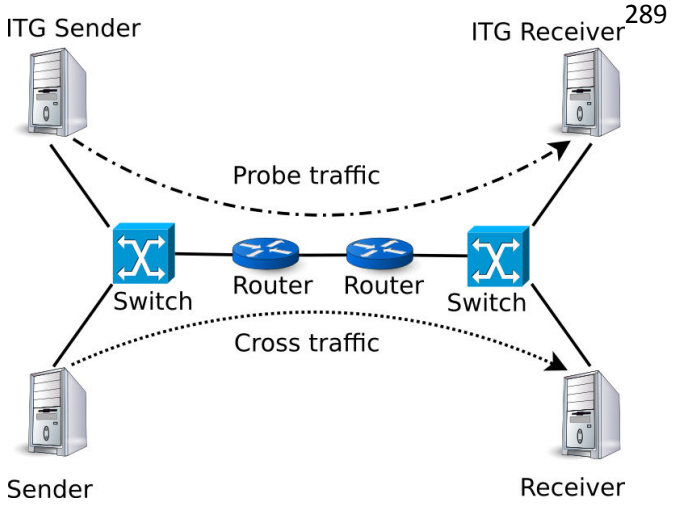

Fig. 5. Testbed configuration.

On the other hand, the length of a REACH only depends on the two parameter σ and γ. In section III-A we calculated the size of a REACH probe as:

$$N_{reach} = 2 \cdot \left\lfloor log_\gamma \left(\frac{\gamma - 1}{\sigma} + 1\right) - 1 \right\rfloor + 2.$$

Hence, our algorithm send always 18 packets using default values of σ and γ.

Our experiments show that the amount of traffic injected by ASSOLO and pathChirp is comparable and extremely limited. Using the default parameters, the measurement process of both tools takes less than one second to produce an estimation over links with a capacity higher than 1 Mbps. However, the two methods are based on the concept of self-induced congestion, i.e., the estimation is performed by injecting probe traffic at a rate higher than the available bandwidth of the network path. The drawback of this approach is that the bottleneck node is congested by the probe traffic – the existing cross-traffic is delayed, and its packets' timings can be significantly affected by the measurement process.

B. Accuracy

We tested both pathChirp and ASSOLO in the presence of different sources of cross-traffic with varying intensity. We generated CBR cross-traffic of 64, 32 and 16 Mbps and, finally, we turned off the traffic source. We evaluated both tools in each cross-traffic scenario, repeating the measurement process 10 times for each algorithm: averaged results are shown in Figure 6. Then, we repeated the same tests simulating different sources of poissonian cross-traffic with increasing average traffic load. The results obtained after 10 runs are shown in Figure 7.

Our experiments show that pathChirp constantly overestimates available bandwidth and measurements are quite unstable. This is a well-know problem of pathChirp: similar results have been obtained in [15], [17], [54]. On the other hand, we found that 80% of ASSOLO's estimations exhibit a relative error lower than 15%. Figure 8 shows an example of

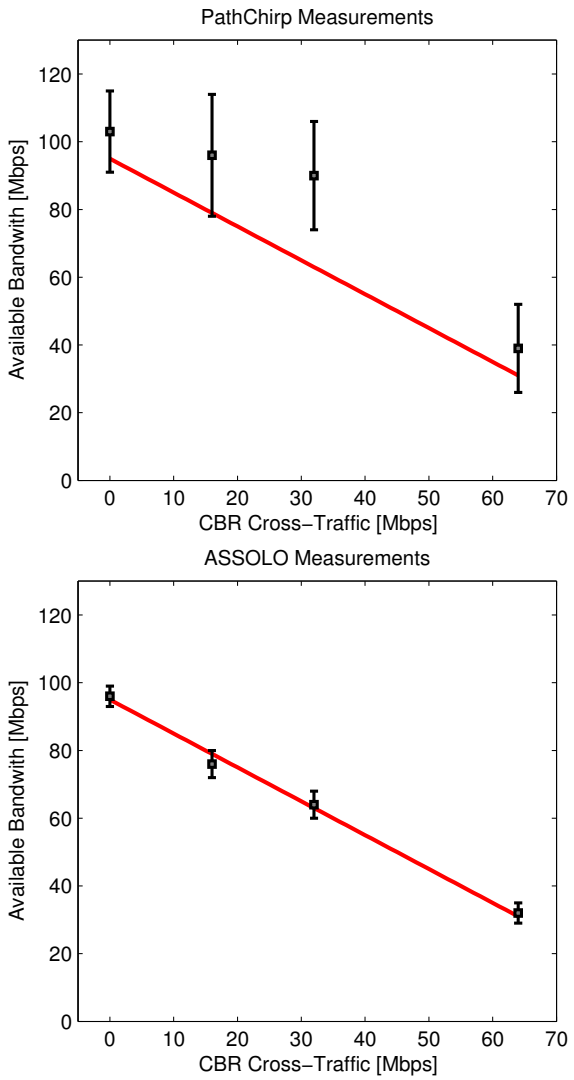

Fig. 6. Measurements obtained in presence of different CBR cross-traffics

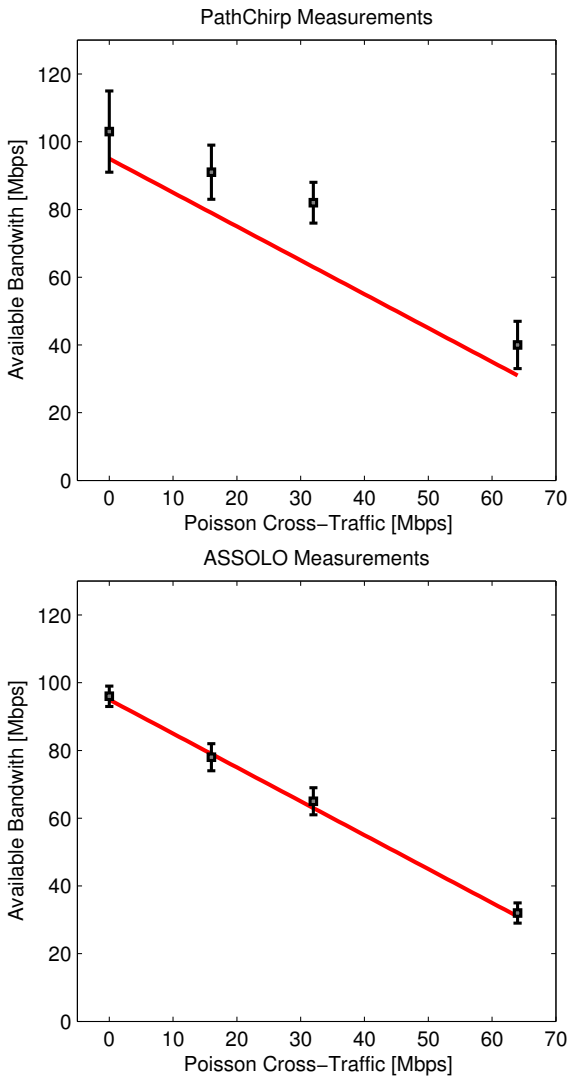

Fig. 7. Measurements obtained in presence of different poissonian cross-traffics

measurement performed in our testbed while the network path is loaded with a Constant Bit Rate cross-traffic of 32 Mbps: the difference between the two tools is notable both in terms of accuracy and stability.

It is worthy of remark that the accuracy of the two tools does not seem to depend on the nature of the cross-traffic – the performances are almost identical using either a CBR source or poissonian distributed packets.

C. Stability

We have analyzed the impact of a real-time operating system on the ASSOLO's measurement process. We performed a few tests with the real-time feature enabled and then we disabled it before repeating the estimation procedure with our tool. A sample comparison of the measurements obtained in the two cases is shown in Figure 9: the average value is correct in both configurations but the real-time feature provides much more stability. Although more investigations would be required, preliminary results confirm that the use of a real-time environment can effectively reduce the impact of different non-deterministic sources of error.

The same experiments could also be repeated for a longer observation interval, in order to catch possible long-term oscillations or biases in the estimations obtained with a non real-time system.

V. Conclusion and Future Work

In this work we presented the details of ASSOLO, an active probing tool which features an efficient measurement scheme for end-to-end available bandwidth estimation in packet-switched networks. Moreover, we described some de-noising techniques and detailed the real-time operating system used by our tool to improve the estimation process.

Preliminary experiments revealed that our algorithm is non-intrusive and accurate, estimating the available bandwidth with greater accuracy and stability with respect to the pathChirp measurement tool developed by the Rice university.

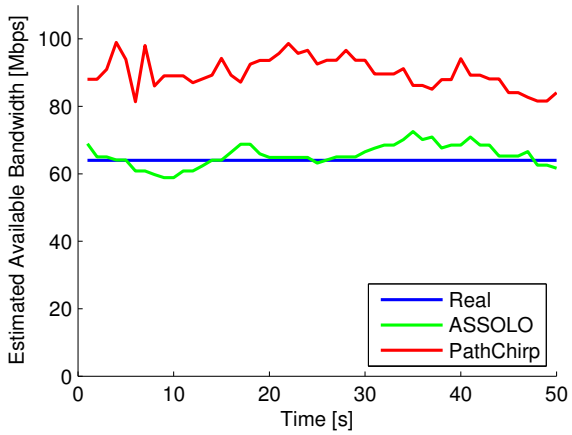

Fig. 8. An example of estimation using pathChirp and ASSOLO.

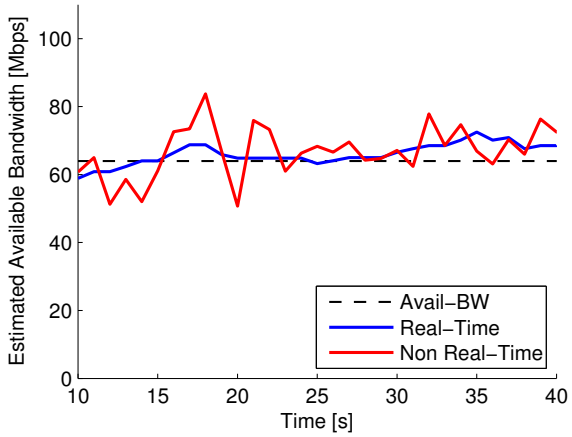

Fig. 9. Measurements using either a real-time operating system or not.

The testbed we used is quite simple and the synthetic cross-traffic does not fully catch the complexity of actual communication flows. We plan to test intensively the performance of our tool over actual Internet paths and in presence of realistic cross-traffic traces. We will also include a study of the actual accuracy, intrusiveness and robustness when dynamic traffic patterns are presents. An extensive comparison of our approach with other state-of-the-art tools is needed too. Above all, BART and FEAT are recent tools which seem to perform better than the original pathChirp: a comparative study will be conducted as soon as the code of these software will be freely available.

Since the bounds of ASSOLO's probing interval have to be set manually at start up, a coarse estimation of the current available bandwidth is required prior using our tool. We plan to introduce an initial self-configuring feature as proposed in [55], thus avoiding the need for any prior knowledge of the network path.

ACKNOWLEDGMENT

We would like to thank Dr. Davide Cavalca for giving the paper a critical reading and for providing us several helpful comments. We acknowledge also Dr. Alberto Savioli and Marco Schivi for their help during the setup of the laboratory testbed and the analysis of experimental results.

REFERENCES

[1] E. Goldoni, G. Rossi, and A. Torelli, "Assolo, a new method for available bandwidth estimation," in *Proc. IARIA International Conference on Internet Monitoring and Protection (ICIMP 2009)*, May 2009, pp. 130–136.

[2] C. Wu, B. Li, and S. Zhao, "Characterizing peer-to-peer streaming flows," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 9, pp. 1612–1626, December 2007.

[3] J. P. Laulajainen, T. Sutinen, and S. Jarvinen, "Experiments with QOS-aware gaming-on-demand service," in *IEEE International Conference on Advanced Information Networking and Applications (AINA 2006)*, vol. 1, Apr. 2006, pp. 805–810.

[4] M. Jain and C. Dovrolis, "Path selection using available bandwidth estimation in overlay-based video streaming," *Computer Networks*, vol. 52, no. 12, pp. 2411–2418, August 2008.

[5] M. Favalli, L.and Folli, A. Lombardo, D. Reforgiato, and G. Schembra, "A bandwidth-aware p2p platform for the transmission of multipoint multiple description video streams," in *Proc. Italian Networking Workshop Reti.it 2009*, Jan. 2009.

[6] T. Tunali and K. Anar, "Adaptive available bandwidth estimation for internet video streaming," *Signal Processing: Image Communication*, vol. 21, no. 3, pp. 217–234, March 2006.

[7] M. Topic, *Streaming Media Demystified*. New York, NY: McGraw-Hill Professional, 2002.

[8] N. Hu and P. Steenkiste, "Evaluation and characterization of available bandwidth probing techniques," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 6, pp. 879–894, August 2003.

[9] J. Strauss, D. Katabi, and F. Kaashoek, "A measurement study of available bandwidth estimation tools," in *Proc. ACM SIGCOMM Conference on Internet Measurement (IMC'03)*, pp. 39–44, October 2003.

[10] M. Jain and C. Dovrolis, "Pathload: A measurement tool for end-to-end available bandwidth," in *Proc. Passive and Active Measurement Conference (PAM 2002)*, Mar. 2002, pp. 14–25.

[11] B. Melander, M. Bjorkman, and P. Gunningberg, "A new end-to-end probing and analysis method for estimating bandwidth bottlenecks," in *Proc. IEEE Global Communications Conference (GLOBECOM 2000)*, Nov. 2000, pp. 415–420.

[12] A. Johnsson, B. Melander, and M. Björkman, "Diettopp: A first implementation and evaluation of a simplified bandwidth measurement method," in *Proc. Swedish National Computer Networking Workshop (SNCNW 2004)*, Nov. 2004.

[13] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell, "pathChirp: Efficient available bandwidth estimation for network paths," in *Proc. Passive and Active Measurement Conference (PAM 2003)*, Apr. 2003.

[14] Q. Wang and L. Cheng, "FEAT: Improving accuracy in end-to-end available bandwidth measurement," in *Proc. IEEE Global Communications Conference (GLOBECOM 2006)*, Nov. 2006, pp. 1–4.

[15] S. Ekelin, M. Nilsson, E. Hartikainen, A. Johnsson, J.-E. Mangs, B. Melander, and M. Bjorkman, "Real-time measurement of end-to-end available bandwidth using kalman filtering," in *Proc. IEEE/IFIP Network Operations and Management Symposium (NOMS 2006)*, Apr. 2006, pp. 73–84.

[16] M. Jain and C. Dovrolis, "Ten fallacies and pitfalls on end-to-end available bandwidth estimation," in *Proc. ACM SIGCOMM Conference on Internet Measurement (IMC'04)*. Oct. 2004, pp. 272–277.

[17] A. A. Ali, F. Michaut, and F. Lepage, "End-to-end available bandwidth measurement tools : A comparative evaluation of performances," in *Proc. International Workshop on Internet Performance, Simulation, Monitoring and Measurement (IPS-MoMe 2006)*, Feb. 2006.

[18] C. D. Guerrero and M. A. Labrador, "On the applicability of available bandwidth estimation techniques and tools," in *Computer Communications*, vol. 33, no. 1, pp. 11–22, January 2010.

[19] L. S. Brakmo, S. W. O'Malley, and L. L. Peterson, "Tcp vegas: new techniques for congestion detection and avoidance," in *Proc. ACM SIGCOMM Conference on Communications Architectures, Protocols and Applications (SIGCOMM'94)*. Aug. 1994, pp. 24–35.

[20] M. Gerla, M. Y. Sanadidi, R. Wang, A. Zanella, C. Casetti, and S. Mascolo, "TCP westwood: congestion window control using bandwidth estimation," in *Proc. IEEE Global Communications Conference (GLOBECOM 2001)*, Nov. 2001, vol. 3, pp. 1698–1702.

[21] C. L. T. Man, G. Hasegawa, and M. Murata, "A new available bandwidth measurement technique for service overlay networks," in *Proc. IEEE/IFIP International Conference on Management of Multimedia Networks and Services (MMNS 2003)*, Sep. 2003, pp. 436–448.

[22] M. Jain and C. Dovrolis, "End-to-end available bandwidth: measurement methodology, dynamics, and relation with tcp throughput," *IEEE/ACM Transactions on Networking*, vol. 11, no. 4, pp. 537–549, August 2003.

[23] L. Lao, C. Dovrolis, and M. Y. Sanadidi, "The probe gap model can underestimate the available bandwidth of multihop paths," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 5, pp. 29–34, October 2006.

[24] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, and R. Baraniuk, "Multifractal cross-traffic estimation," in *Proc. ITC Specialist Seminar on IP Traffic Measurement*, Sep. 2000.

[25] A. Pasztor, "Accurate active measurement in the internet and its applications," Ph.D. dissertation, Department of Electrical and Electronic Engineering, The University of Melbourne, 2003.

[26] H. Zhou, Y. Wang, X. Wang, and X. Huai, "Difficulties in estimating available bandwidth," in *Proc. IEEE International Conference on Communications (ICC'06)*, Jun. 2006, pp. 704–709

[27] The Linux Kernel. [Online]. Available: http://www.kernel.org

[28] Y. Ozturk and M. Kulkarni, "Dichirp: direct injection bandwidth estimation," *International Journal of Network Management*, vol. 18, no. 5, pp. 377–394, September 2008.

[29] Gnu/Linux Real-Time. [Online] Available: http://rt.wiki.kernel.org/

[30] K. Koolwal, "Myths and realities of real-time linux software systems," in *Proc. Real-Time Linux Workshop (RTLWS 2009)*, Oct. 2009. [Online]. Available: http://lwn.net/images/conf/rtlws11/papers/proc/p20.pdf

[31] RTAI: Realtime application interface for linux. [Online] Available: http://www.rtai.org/

[32] Xenomai: Real-time framework for Linux. [Online] Available: http://www.xenomai.org

[33] K. Yaghmour, J. Masters, G. Ben-Yossef, and P. Gerum, *Building Embedded Linux Systems*. Sebastopol, CA: O'Reilly & Associates, 2008.

[34] R. Prasad, M. Murray, C. Dovrolis, and K. Claffy, "Bandwidth estimation: Metrics, measurement techniques, and tools," IEEE Network, vol. 17, no. 6, pp. 27–35, November 2003.

[35] C. D. Guerrero and M. A. Labrador, "Experimental and analytical evaluation of available bandwidth estimation tools," in *Proc. IEEE Conference on Local Computer Networks (LCN 2006)*, Nov. 2006, pp. 710–717.

[36] E. Goldoni, "Nuovi approcci nella stima della banda disponibile in una rete a pacchetto," Master thesis, University of Pavia, 2007.

[37] J. Navratil and R. L. Cottrell, "Abwe: A practical approach to available bandwidth," in *Proc. Passive and Active Measurement Conference (PAM 2003)*, Apr. 2003.

[38] J. Sommers, P. Barford, and W. Willinger, "A proposed framework for calibration of available bandwidth estimation tools," in *Proc. IEEE Symposium on Computers and Communications (ISCC'06)*, Jun. 2006, pp. 709–718.

[39] S.-R. Kang and D. Loguinov, "IMR-Pathload: Robust available bandwidth estimation under end-host interrupt delay," in *Proc. Passive and Active Measurement Conference (PAM 2008)*, Apr. 2008, pp. 172–181.

[40] G. Urvoy-Keller, T. En-Najjary, and A. Sorniotti, "Operational comparison of available bandwidth estimation tools," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 1, pp. 39–42, January 2008.

[41] Y. Zhang and N. Duffield, "On the constancy of Internet path properties," in *Proc. ACM SIGCOMM Workshop on Internet Measurement (IMW'01)*, Nov. 2001, pp. 197-211.

[42] L. Burgstahler and M. Neubauer, "New modifications of the exponential moving average algorithm for bandwidth estimation," in *Proc. ITC Specialist Seminar on Internet Traffic Engineering and Traffic Management*, July 2002.

[43] M. Kim and B. Noble, "Mobile network estimation," in *Proc. International conference on Mobile Computing and networking (MobiCom'01)*, Jul. 2001, pp. 298–309.

[44] E. Goldoni, G. F. Rossi, and P. Gamba, "Improving available bandwidth estimation using averaging filtering techniques," University of Pavia, Tech. Rep., 2008. [Online]. Available: netlab-mn.unipv.it/publications/tr-netlab2008-01.pdf

[45] A. White, "The vertical horizontal filter," *Futures Magazine*, vol. 20, no. 10, pp. 1–10, 1991.

[46] J. C. Mogul and K. K. Ramakrishnan, "Eliminating receive livelock in an interrupt-driven kernel," *ACM Transactions on Computer Systems*, vol. 15, no. 3, pp. 217–252, August 1997.

[47] Intel. (2003) Interrupt moderation using Intel gigabit ethernet controllers. [Online]. Available: http://download.intel.com/design/network/applnots/ap450.pdf

[48] J. H. Salim, R. Olsson, and A. Kuznetsov, "Beyond softnet," in *Proc. USENIX Annual Linux Showcase & Conference (ALC'01)*. Nov. 2001, pp. 165-172.

[49] R. Prasad, M. Jain, and C. Dovrolis, "Effects of interrupt coalescence on network measurements," in *Proc. Passive and Active Measurement Conference (PAM 2004)*, Apr. 2004, pp. 247–256.

[50] S. Avallone, S. Guadagno, D. Emma, A. Pescapè, and G. Ventre, "D-ITG distributed internet traffic generator." in *QEST*. IEEE Computer Society, 2004, pp. 316–317.
A. Botta, A. Dainotti, A. Pescapè, "Multi-protocol and multi-platform traffic generation and measurement," in *IEEE Conference on Computer Communications (INFOCOM 2007), Demo Session*, May 2007.

[51] R. S. Prasad, C. Dovrolis, and B. A. Mah, "The effect of layer-2 store-and-forward devices on per-hop capacity estimation," in *IEEE Conference on Computer Communications (INFOCOM 2003)*, Mar. 2003, vol. 3, pp. 2090–2100.

[52] L. Angrisani, S. D'Antonio, M. Esposito, and M. Vadursi, "Techniques for available bandwidth measurement in ip networks: a performance comparison," *Computer Networks*, vol. 50, no. 3, pp. 332–349, February 2006.

[53] A. Torelli, "Sviluppo di una tecnica innovativa per la stima della banda disponibile," Master thesis, University of Pavia, 2008.

[54] A. Shriram, M. Murray, Y. Hyun, N. Brownlee, A. Broido, and K. C. M. Fomenkov, "Comparison of public end-to-end bandwidth estimation tools on high-speed links," in *Proc. Passive and Active Measurement Conference (PAM 2005)*, Mar. 2005, pp. 306–320

[55] W. Tan, M. Zhanikeev, and Y. Tanaka, "Abshoot: A reliable and efficient scheme for end-to-end available bandwidth measurement," in *Proc. IEEE Region 10 Conference (TENCON 2006)*, Nov. 2006.

www.iariajournals.org

International Journal On Advances in Intelligent Systems
ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS
issn: 1942-2679

International Journal On Advances in Internet Technology
ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING
issn: 1942-2652

International Journal On Advances in Life Sciences
eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO
issn: 1942-2660

International Journal On Advances in Networks and Services
ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION
issn: 1942-2644

International Journal On Advances in Security
ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS
issn: 1942-2636

International Journal On Advances in Software
ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS
issn: 1942-2628

International Journal On Advances in Systems and Measurements
ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL
issn: 1942-261x

International Journal On Advances in Telecommunications
AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA
issn: 1942-2601