Economics of CPU in DRAM and TOMI™ Technology

by Kendall Laughlin and Russell Fish III, Venray Technology Ltd

The Year an Industry Died

Minicomputers died in 1974, murdered by microprocessors. The body twitched for another decade, but it was still a TTL laden, Unibus connected, paper tape consuming corpse.

- Minicomputers didn't die because microprocessors were faster. They weren't
- Minicomputers didn't die because microprocessors had more software. They did not.
- Minicomputers didn't die because microprocessors were easier to use. They were not.

Minicomputers died because microprocessors collapsed the cost of computing power and with it the business model for a multi-billion dollar industry.

In 1974 Digital Equipment's entry level minicomputer, the PDP-8, cost $15,000. In 1974 both Intel and Motorola introduced microprocessors that sold for $300, and in 1975 MOS Technology announced the 6502 for $25.

PDP 8

However microprocessors didn't just crush minicomputers with cheap chips. By dramatically lowering the cost of computing they conjured previously inconceivable new industries out of thin air including cell phones, video games, music players, and of course PCs.

It's 1974 Again

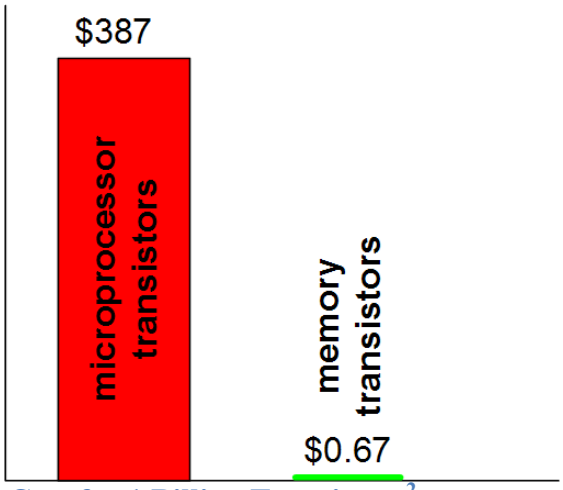
While both the computer industry and government have been moaning "Game Over"[1] due to excessive microprocessor heat and memory congestion, costs are preparing to replay 1974.

Computing power is about to get a whole lot cheaper.

---

[1] Exascale Computing Study. DARPA http://users.ece.gatech.edu/~mrichard/ExascaleComputingStudyReports/exascale_final_report_100208.pdf and The Future of Computing Performance, Game Over or Next Level?, National Research Council http://sites.nationalacademies.org/CSTB/CurrentProjects/CSTB_042221

Illustrating the changing economics:



Cost for 1 Billion Transistors[2]

Both microprocessors and memory chips are constructed from billions of electrical switches called transistors. Microprocessor transistors are manufactured using one recipe. Memory transistors are manufactured using another.

	Microprocessor Transistor	Memory Transistor
Speed	1x	0.8x
Leaked Power	1x	0.00001x
Cost	1x	0.01x

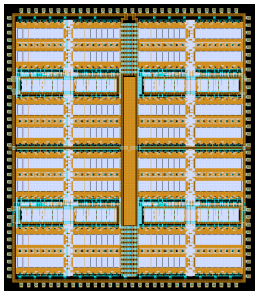
Comparing Transistor Parameters

Microprocessor transistors are built for speed and are somewhat faster than similar memory transistors. Memory transistors are built for low power leakage because leakage makes memories forget. Memory transistors are simpler to build since they are connected in regular patterns and need many fewer layers of wires. Connecting wires are very expensive to make. As a result, memory transistors are really really cheap.

Merging Microprocessor and Memory

Memory transistors can perform the same functions as microprocessor transistors. This means microprocessors can be designed using cheap low leakage memory transistors.

Furthermore the microprocessors designed with memory transistors can be combined on the same chip with the memories themselves.



TOMI Aurora

[2]$387 from INTEL Xeon E5620, http://ark.intel.com/products/family/28144
$0.67 from DRAM Exchange, October 3, 2011, http://www.dramexchange.com

This significantly speeds up connections between the two and lowers cost even further.

The idea of building microprocessors on memory chips was first proposed over two decades ago[3], but until recently, no designs were done using the technique. When first suggested, INTEL dominated the microprocessor world with its "industry standard" architecture and it was considered suicidal to compete against their market power and alliances with Microsoft. In the 90s, the government funded about a dozen designs to increase computer performance by building memories using logic transistors[4], but the chips were predictably very expensive.

INTEL's near total dominance discouraged new designs until recently. Three changes broke INTELs market hold:

- INTEL lost its software monopoly.
- INTEL chips ran too hot to sell.
- INTEL multi-core turned out to be a technical bust.

Bested by Bohemians

In 1981, IBM introduced the Personal Computer featuring an INTEL microprocessor. At that time there were other computers using other microprocessors, but programs written for the PC could only run on the INTEL microprocessor. The PC was a huge success and swept away almost all competitors and their microprocessor suppliers. Most programs were written for the PC and could therefore only run on INTEL microprocessors.

In the late 80s Richard Stallman[5] with the eventual assistance of Michael Tiemann[6] began creating a program that could enable other programs to run on just about any microprocessor whether INTEL or not. In 1991 a Finnish student, Linus Torvalds, began writing a program to manage the most basic operations of a computer. Together these two programs became Linux[7], the tool that would eventually break the monopoly of INTEL microprocessors and Microsoft software.

Linux made it possible to write one program to run on many microprocessors, including those not made by INTEL.

---

[3] Fig 9, US Patent 5,440,749, filed Aug, 3, 1989
http://www.pat2pdf.org/pat2pdf/foo.pl?number=5,440,749
[4] Intelligent RAM, http://www.eecs.berkeley.edu/~pattrsn/talks/iram.html; Execube PIM, http://www.nd.edu/~pim/projects.html; Cyclops, http://www.research.ibm.com/people/c/cascaval/medea02.pdf
[5] http://en.wikipedia.org/wiki/Richard_Stallman
[6] http://en.wikipedia.org/wiki/Michael_Tiemann
[7] Linux, http://en.wikipedia.org/wiki/Linux

Coal Powered Microprocessors Run Out of Steam

The industry was stunned when INTEL cancelled its two premier processor designs in May of 2004[8]. They immediately spun the failure as a strategic redirection, but outside analysis pointed to a more sinister problem. INTEL's Tejas, Sanscrit for fire[9], dissipated a stupendous 150 watts at 2.8 GHz[10], more than Hasbro's Easy Bake Oven[11].

The Tejas had been projected to run 7 GHz[12]. It never did, and no INTEL processor ever has. When microprocessors run faster they get hotter. When they get too hot, they quit working and sometimes blow up[13]. INTEL wisely chose to build many slower processors instead of one fast hot one. They called the technique multi-core which meant that two or more complete microprocessors were built on the same chip and attached to a shared memory bus.

Multi-core Epic Fail

Multi-core was introduced as a marketing strategy to justify higher prices to customers purchasing slower microprocessors. One core must be good, so two must be twice as good, and 4 must be 4 times, even if the clock speed was not so much.

The situation was similar to the transistor radio wars of the 1960s[14]. The Japanese had dominated the radio business ever since Sony acquired a license to manufacture transistors from Bell Labs in 1954[15]. Ten years later the technology had matured and Hong Kong began producing radios. For a decade the standard handheld radio had been based on a 6 transistor design that traced back to Texas Instruments. Sometimes a manufacturer would add an additional amplifier and the design would be 7 or 8 transistors.

However in order to differentiate themselves from the Japanese, the Hong Kong competitors began "adding" transistors. They did not add electronic circuitry that used these transistors and improved performance. They simply "added" the transistors to the circuit board[16]. Often all 3 transistor leads were tied together and connected to ground. Many were just dummies. Within a year the standard radio went from 6 transistors, to 8,

[8] "INTEL Halts Development of 2 New Processors", NY Times http://www.nytimes.com/2004/05/08/business/08chip.html?ex=1399348800&en=98cc44ca97b1a562&ei=5007

[9] Tejas, a Sanskrit synonym for fire, http://en.wikipedia.org/wiki/Tejas

[10] Intel Tejas Samples Dissipate 150W of Heat at 2.80 GHz http://www.xbitlabs.com/news/cpu/display/20040111115528.html

[11] Easy Bake Oven, http://en.wikipedia.org/wiki/Easy-Bake_Oven

[12] Intel Unveils Tejas Platform Details at IDF, 7 GHz, http://www.xbitlabs.com/news/cpu/display/1045735111.html

[13] http://www.youtube.com/watch?v=wSALep8QZ84

[14] Phony Transistor Counts, http://www.monitoringtimes.com/MT/lowressample.pdf

[15] http://www.sony.net/SonyInfo/CorporateInfo/History/SonyHistory/index.html

[16] http://www.flickr.com/photos/erinbanister/751655481/

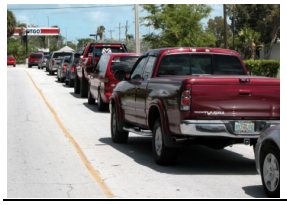
to 10, and eventually to 16. Eventually customers wised up to the scam and the portable electronics business went into decline not reversed until the arrival of the Sony Walkman.

Multi-core microprocessors are different from the dummy transistor radios because the additional microprocessor cores consist of operating transistors that are actually connected to circuits and actually affect performance. **Mostly they consume power and generate heat**. The multiple cores exist, but they are prevented from doing much useful work.

In small amounts multi-core can be effective. Two or even 4-cores can improve performance. However, doubling the number of cores does not double the performance. But the real news is even worse for popular cloud computing applications such as managing unstructured data. Sandia Labs performed an analysis of multi-core microprocessors running just such "data intensive" applications.

Sandia Labs[17] is an 8,000 person US government lab that dates back to the Manhattan Project. Besides nuclear weapons work, Sandia hosts several supercomputers and is a major computer research center.

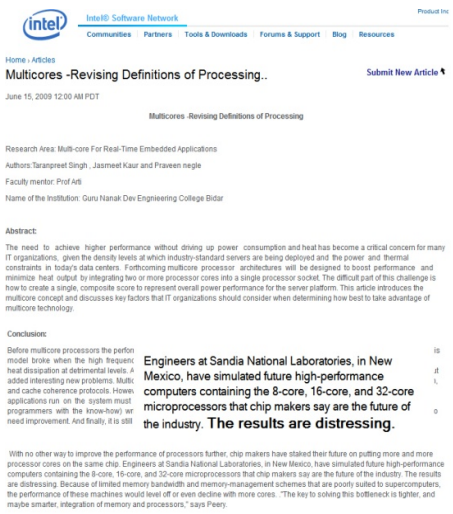
From their release, "A Sandia team simulated key algorithms for deriving knowledge from large data sets. The simulations show a significant increase in speed going from two to four multicores, but an insignificant increase from four to eight multicores." Sandia Labs found that for data-intensive programs **16-core microprocessors delivered the same performance as 2-cores**[18].



Sandia described the bottleneck thusly, "The problem is the lack of memory bandwidth as well as contention between processors over the memory bus available to each processor." Like waiting for the gas pump.

[17] Sandia Labs, http://en.wikipedia.org/wiki/Sandia_National_Laboratories
[18] https://share.sandia.gov/news/resources/news_releases/more-chip-cores-can-mean-slower-supercomputing-sandia-simulation-shows/

INTEL has acknowledged the problem as "**a critical concern**"[19] and recognized the criticality with the following understatement, "Engineers at Sandia National Laboratories, in New Mexico, have simulated future high-performance computers containing the 8-core, 16-core, and 32-core microprocessors that chip makers say are the future of the industry. **The results are distressing**."


Intel® Software Network

Communities | Partners | Tools & Downloads | Forums & Support | Blog | Resources

Home › Articles

Multicores -Revising Definitions of Processing..

Submit New Article

June 15, 2009 12:00 AM PDT

Multicores -Revising Definitions of Processing

Research Area: Multi-core For Real-Time Embedded Applications

Authors:Taranpreet Singh , Jasmeet Kaur and Praveen negle

Faculty mentor: Prof Arti

Name of the Institution: Guru Nanak Dev Engnieering College Bidar

Abstract:

The need to achieve higher performance without driving up power consumption and heat has become a critical concern for many IT organizations, given the density levels at which industry-standard servers are being deployed and the power and thermal constraints in today's data centers. Forthcoming multicore processor architectures will be designed to boost performance and minimize heat output by integrating two or more processor cores into a single processor socket. The difficult part of this challenge is how to create a single, composite score to represent overall power performance for the server platform. This article introduces the multicore concept and discusses key factors that IT organizations should consider when determining how best to take advantage of multicore technology.

Conclusion:

Before multicore processors the perfor model broke when the high frequenc heat dissipation at detrimental levels. A added interesting new problems. Multic and cache coherence protocols. Howev applications run on the system must programmers with the know-how) wr need improvement. And finally, it is still

Engineers at Sandia National Laboratories, in New Mexico, have simulated future high-performance computers containing the 8-core, 16-core, and 32-core microprocessors that chip makers say are the future of the industry. The results are distressing.

With no other way to improve the performance of processors further, chip makers have staked their future on putting more and more processor cores on the same chip. Engineers at Sandia National Laboratories, in New Mexico, have simulated future high-performance computers containing the 8-core, 16-core, and 32-core microprocessors that chip makers say are the future of the industry. The results are distressing. Because of limited memory bandwidth and memory-management schemes that are poorly suited to supercomputers, the performance of these machines would level off or even decline with more cores. "The key to solving this bottleneck is tighter, and maybe smarter, integration of memory and processors," says Peery.

Hail Mary

Superstar chip architect, David Patterson of Berkeley, is slightly more harsh,[20] describing

INTEL's multi-core strategy as, "the equivalent of a Hail Mary pass when it switched from making microprocessors run faster to putting more of them on a chip." He elaborated, ".....a desperate pass with little chance of success has been labeled a Hail Mary." Not only is Patterson critical of multi-core power, memory bandwidth, and parallelism, he also cites programming challenges.

INTEL's only response so far appears to be their rendition of "Back to the Future"[21]. They call it, Hybrid Memory Cube[22]. We call it "Rambus Redux"[23].

TOMI Technology and TOMI Borealis

When INTEL was finally vulnerable, TOMI engineers got busy.

To create TOMI Technology they had to solve two large problems that had prevented others from designing microprocessors using memory transistors:

1. Microprocessors are created from collections of pre-defined modules. Each module consists of many transistors wired together to perform a specific function such as ADD. No such libraries of modules existed for memory transistors.
2. Memory transistors have very few wires with which to make connections. Microprocessors usually contain tens of millions of transistors wired together with many wires.

---

[19] http://software.intel.com/en-us/articles/multicores-revising-definations-of-processing/

[20] "The Trouble with Multi-core", David Patterson http://spectrum.ieee.org/computing/software/the-trouble-with-multicore/0

[21] "Back to the Future". http://www.imdb.com/title/tt0088763/

[22] "Reinventing DRAM with the Hybrid Memory Cube", http://blogs.intel.com/research/2011/09/hmc.php

[23] Rambo Rambus.(Intel and Rambus develop new faster dynamic random-access memory computer chip) http://www.highbeam.com/doc/1G1-19104133.html

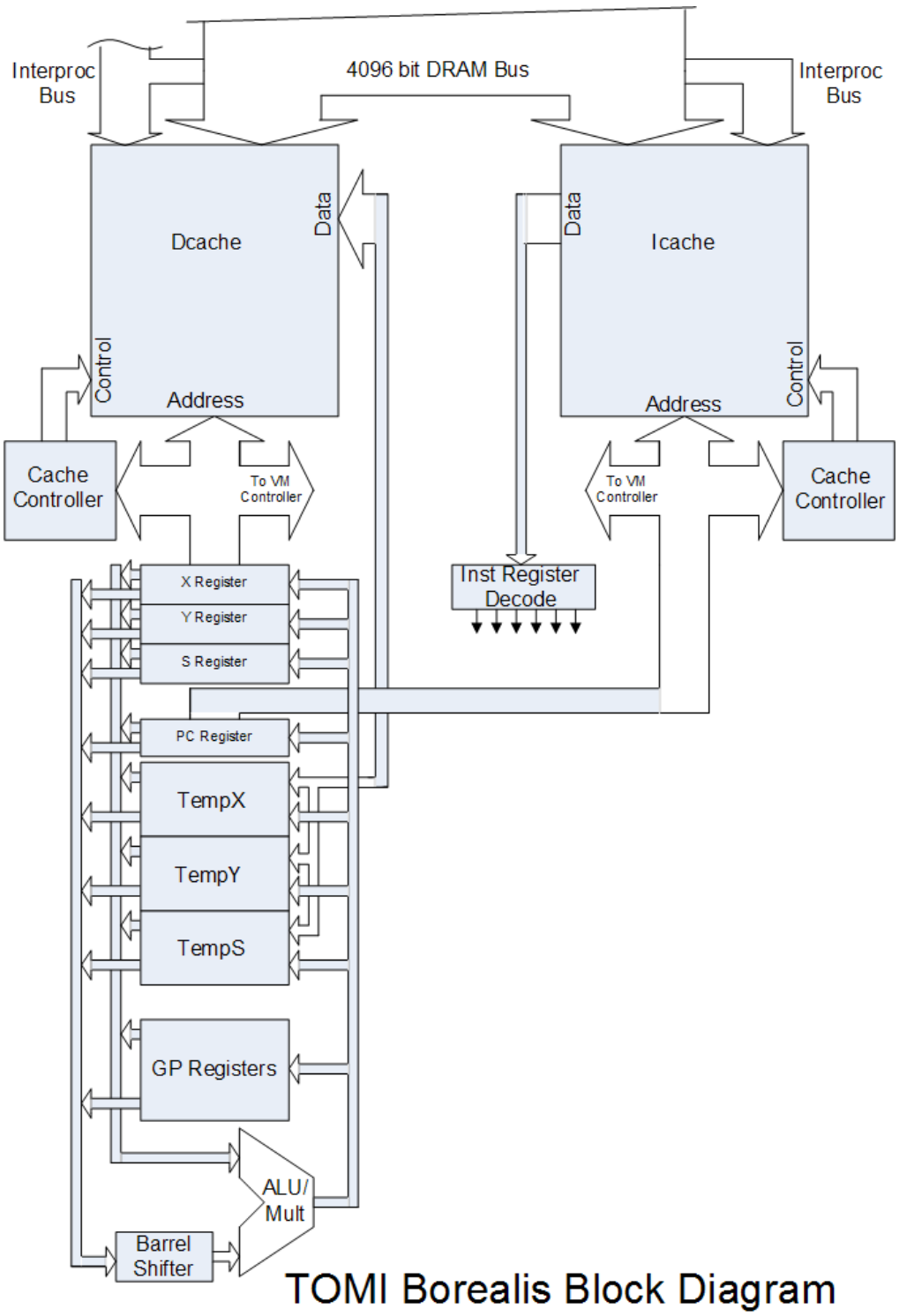
The TOMI designers solved the module problem by creating and characterizing libraries of both digital and analog modules for several memory transistor recipes. "Characterizing" means that the performance of each module was documented over a range of voltage and temperature conditions. These characterized modules would be wired together to make the TOMI microprocessors.

Memory transistors do not have enough wires to connect the tens of millions needed for a legacy microprocessor such as and INTEL Xeon or AMD Phenom. In order to make a microprocessor using memory transistors TOMI designers had to invent a very simple architecture that required few transistors yet performed well on popular programs, particularly those handling unstructured data[24] in Cloud Computing[25].

The resulting TOMI Borealis design shown in Fig. 4 includes eight 32-bit microprocessors embedded in a commodity 1G DRAM chip. The architecture can be considered a modified RISC[26] (Reduced Instruction Set Computer). Each microprocessor core, less caches, requires fewer than 22,000 transistors including multiplier and barrel shifter. In order to dramatically reduce the transistor count to fit the available connecting wires, each design choice had its performance effect weighed against its cost in transistors.

---

[24] http://en.wikipedia.org/wiki/Unstructured_data
[25] http://en.wikipedia.org/wiki/Cloud_computing
[26] http://en.wikipedia.org/wiki/Reduced_instruction_set_computing

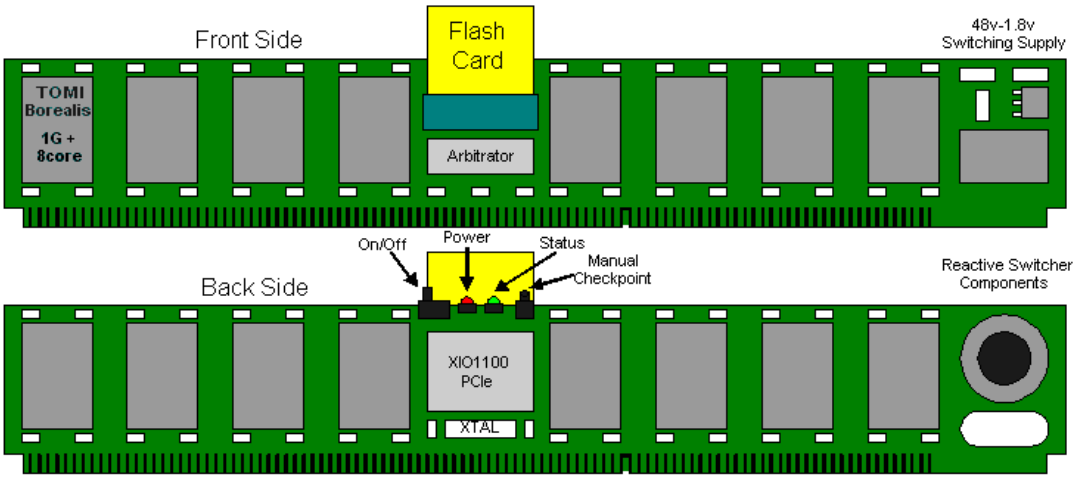


TOMI Borealis Block Diagram

The wiring was simplified by creating each of the 32-bits individually and then connecting the bits together with control lines. (The technique is called bit-slicing.) As an added advantage, bit-slicing greatly eases moving the microprocessor from one DRAM design to another.

Of course TOMI Borealis does not suffer from INTEL's crippling inability to bring memory data to its multicores. TOMI cores reside in the middle of memory. **The data is already there**.

Sixteen TOMI Borealis chips were configured to fit on a 4 inch circuit board about the size of an ordinary memory DIMM. The board includes 128 cores, 2G of DRAM, network controller, and a switching power supply. It is shown below in Fig. 5 with the heat spreader removed.



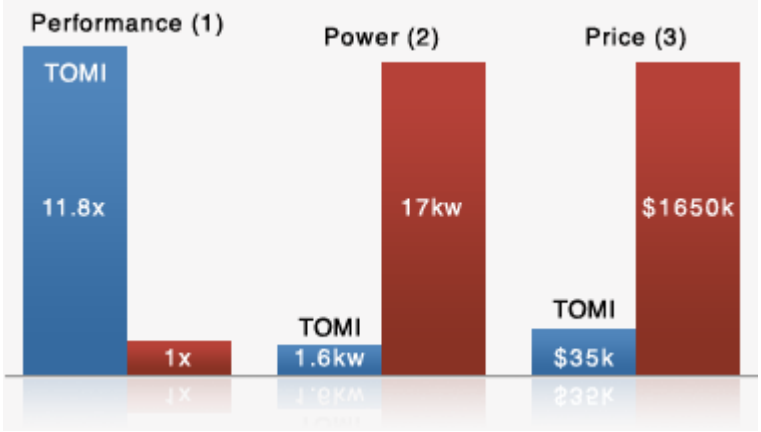
For massively parallel benchmarking purposes, we assumed four rows of 32 of these DIMMs arranged on a single 19" motherboard with provisions for forced air cooling. This was compared against a full 19" rack of INTEL Xeon processors.

Fig. 6 illustrates the results comparing the 2.1 GHz TOMI Borealis system and the INTEL's 2.4 GHz Xeon E5620 system running Sandia Labs MapReduce-MPI[27]. MapReduce is one of the most commercially popular massively parallel applications. It divides huge databases across many microprocessors in order to perform rapid searches in parallel. MapReduce was first proposed by Google to coordinate the million servers that constitute its search engine, probably the largest cloud computing installation in existence[28].

MapReduce-MPI is written in C++ and tuned for Sandia's supercomputers, such as Red Storm[29]. It is therefore much faster than Hadoop, the popular MapReduce implementation written in Java[30]. The results of the comparison are found below.

[27] Sandia Labs, MapReduce-MPI, http://www.sandia.gov/~sjplimp/mapreduce.html
[28] http://en.wikipedia.org/wiki/Google
[29] https://share.sandia.gov/news/resources/news_releases/national-security-computing-center-open-for-business/
[30] http://hadoop.apache.org/

TOMI™ Technology and the Cloud



TOMI Borealis vs. Legacy XEON E5620 MPP

(1) Performance
Sandia Labs MapReduce, 256 GB dataset, 10/1 mix retrieve/insert unstructured data.

(2) Power
Per Oracle datasheet 173705.pdf, per Venray Technology Ltd 42nm TOMI Borealis.

(3) Price
Per Oracle Exadata Price List 070598.pdf, per Venray Technology Ltd, 65% gross margin, Nanya, Samsung, or Promos fab.

Implications of TOMI Technology

Productivity is what makes economies grow[31]. Improved productivity means an economy receives more goods and services for the same labor and capital. Computers significantly increase economic productivity by allowing people to get more work done[32]. The National Research Council has warned "Game Over" for productivity if improvements in computing power are stopped due to power and memory problems (2).

Since TOMI Technology significantly reduces the cost of computing power while simultaneously mitigating other problems including both power and memory congestion, it will increase productivity. Some areas of improvement are clearly visible now, some only future shadows.

Initially TOMI Technology will make cloud servers significantly less expensive, more powerful, and more energy efficient.

With very little difficulty a single one of the 8-core TOMI Borealis chips that implements a massively parallel cloud server could significantly reduce the power consumption of battery powered devices such as tablets and cell phones. Each one is less than 20% larger than the 65 cent 1GB DRAM on which it is based.

[31] http://en.wikipedia.org/wiki/Productivity
[32] Computing Productivity, http://web.mit.edu/sloan-msa/Papers/2.6.pdf

However just as with the 1974 microprocessor introductions, the greatest productivity will come from classes of products not previously possible. In 1974 we squinted into the future trying to glimpse the microprocessor products to come. None of us at Motorola could see video games, cell phones, and certainly not the PC.

"*It's tough to make predictions, especially about the future.*" - Yogi Berra[33]

The innovators such as Chuck Peddle[34], Nolan Bushnell[35], Steve Jobs[36], Steve Wozniak[37], and Martin Cooper[38] created the future that the early microprocessor architects could only enable.

Today in 2011 we can predict TOMI will accelerate some existing trends, but entrepreneurs and innovators will fill in the specifics. Other than one obvious implication we can only hypothesize about the future. Predicting the future is the rarefied area of masters such as Kurtzweil[39] and of course Jobs.

Obviously the INTEL multi-core architecture is as dead in 2011 as the PDP-8 was in 1974. Like the PDP-8, the corpse will continue to twitch for a decade before eventually retiring to a position of honor[40].

Through the crystal ball we can only perceive the faint outlines of what new might come. Here goes, including both the opportunities and the dangers:

- Continuous speech recognition will finally be embedded in every product that interfaces with humans. Every light switch and every appliance will be voice controlled, and some will talk back.
- Computer speech will be so common that devices will talk to other devices in the presence of their users.
- Common commands will be reduced to short verbal cues like clicking your tongue or sucking your lips.
- Language translation will be built into every cell phone.
- Your cell phone will identify who is speaking by their voice.

[33] Yogi Berra, http://quotationsbook.com/quote/16478/
[34] http://en.wikipedia.org/wiki/Chuck_Peddle
[35] http://en.wikipedia.org/wiki/Nolan_Bushnell
[36] http://en.wikipedia.org/wiki/Steve_Jobs
[37] http://en.wikipedia.org/wiki/Steve_Wozniak
[38] http://en.wikipedia.org/wiki/Martin_Cooper_(inventor)
[39] http://en.wikipedia.org/wiki/Ray_Kurzweil
[40] http://www.computerhistory.org/

- Pharmaceutical companies will give away devices that play free music while constantly monitoring medical parameters and from time to time suggesting products, drugs, treatments, or exercise regimens.
- Insurance companies will give discounts to customers who use similar devices that not only monitor health, but predict imminent problems, and inform the user as well as the company.
- Your cell phone camera will recognize the face of anyone it sees and scan the computer cloud for background red flags as well as "Six degrees of separation"[41].
- TOMI enabled devices will be light enough to build into glasses with a camera. Everywhere the user glances, the camera will see faces, buildings, UPC codes, or license plates, search the cloud, and highlight the item with specific details as the viewer looks through the translucent LCD glasses. The device will also play media on the same viewer.

- Every retailer will give away these glasses for the right to collect your purchasing habits. Retailers may group together into non-competitive information sharing cooperatives. When you walk into a physical retailer, your presence will immediately be known as well your habits and purchasing history. From the glasses, you may be offered sale specials or suggestions consistent with previous habits.
- Every large city will become like a small town. Your personal history both good and bad will be available for one and all to see, as your identity is displayed by your face and confirmed by your voice. Credit bureaus will go out of business. You will never have to present an ID. Women will create search engines to find eligible and prosperous men. Men will create search engines to qualify women. Your resume and criminal record will follow you everywhere.
- Drivers will become much friendlier since everyone will know the identity of both the driver in front and the one behind. Drivers with many infractions can immediately be identified from the wide berth given by other motorists.
- Criminals will find their jobs much more difficult because their history will be immediately visible to anyone who encounters them.
- TOMI Technology will be built on flash memories creating the elemental unit of a learning machine. Initial applications will perform small mundane tasks not much smarter than Roomba[42]. However the machines will be able to self organize, build robust communicating structures, and collaborate to perform tasks. Each machine will have knowledge of its mechanical capabilities and be programmed with a series of goal-seeking behaviors and simple strategies. Each strategy will be tried to satisfy the goal, the results memorized in flash, and used to modify the strategies. The first goal will be to seek power when it runs low. Machines will

[41] http://en.wikipedia.org/wiki/Six_degrees_of_separation
[42] http://en.wikipedia.org/wiki/Roomba

be small, weigh a pound or less, and since they will cost less than $20, they will be expendable. Tasks will range from cutting grass, to picking up trash, to paving a street, to painting a building, to clearing a minefield.

- The same intelligence used in the self-teaching machines will crawl the terabytes of unstructured enterprise data looking for patterns and relationships. High-velocity management will become more valuable than high-velocity trading.
- Cloud computing nodes will be so inexpensive and redundant that 100,000 cores will fit in a child's lunch box. Cloud providers will pay struggling cable television companies to place these mini-clouds in their switching boxes where high speed Internet connectivity and power already exist.
- Cloud providers will pay cash strapped school districts to connect 2 million-core TOMI Technology vending machine sized servers to their football stadium power grids. The containers will be online except during games.
- A disposable diaper company will give away TOMI enabled teddy bears that teach reading and arithmetic. It will be able to identify specific children and complain when it is tired (needs recharging). The bear will play a game upon the successful completion of a lesson, and from time to time remind Mom to buy a product. The bear will also diagnose a raspy throat, a cough, or runny nose and suggest another product. Due to the high commercial value of being able to analyze and recommend products directly to customers, there will be fierce competition among diaper companies to offer the most effective reading teaching system in their bear.

It will be a brave new world, and TOMI Technology will enable it.

Kendall Laughlin

Kendall has spent 30 years in consumer and industrial marketing. He is currently Director of International Sales for Venray Technology Ltd, an IP development concern. He has a BS in Economics from the University of Arkansas and a JD from the SMU School of Law.

Russell Fish III

Russell's three-decade career dates from the birth of the microprocessor. One or more of his designs are licensed into most computers, cell phones, and video games manufactured today.

Russell and Chuck Moore created the Sh-Boom Processor which was included in the IEEE's "25 Microchips That Shook The World". He has a BSEE from Georgia Tech and an MSEE from Arizona State.