

Moore's Law and Cluster Computing: When Moore Is Not Enough

Greg Lindahl
Distinguished Engineer
Key Research, Inc.

Cluster Computing 101

“Cluster computing is easy. Why, you just throw some stuff together, and it works. And best of all, it all just gets better and better, because computers get faster because of Moore's Law!”

Talk Outline

- What is Moore's Law?
- Why you should care
- Rapidly improving technologies
- Slowly improving technologies
- Gut-wrenching Nightmares
- Conclusion

What is Moore's Law, anyway?

- Gordon Moore of Intel Corp.
- First stated in 1965: the number of transistors on a chip will double each year, for the next 10 years.
- Updated in 1975: the number of transistors on a chip will double each 2 years.
- Widely misquoted with the time constant changed to 18 months.

What it is NOT about:

- Performance
- Feature size
- Transistors / area
- Price
- Power
- Clock
- Bandwidth
- Disks
- Lasers / optics
- Wires
- It's not 18 months.
- It's not about the speed of light.
- It's not about latency.

The pundit speaks:

"Moore's Law has been the name given to everything that changes exponentially in the industry. I saw[sic], if Gore invented the Internet, I invented the exponential." --Gordon Moore

<http://www.usnews.com/usnews/transcripts/moore.htm>

Sometimes it is true...

- Some things really do improve with Moore's Law
- Example: CCDs used in digital cameras are built with semiconductor process similar to CMOS
- So you can chart the demise of film.
- The production of photographic plates for astronomy died out ~ 1995!

But:

"Rumors of my demise are greatly exaggerated."

-- Mark Twain (maybe)

- Moore's law has become a self-fulfilling prophecy: Semiconductor companies use it to guide their R&D budgets.
- Improvements in clock / size / process / memory / compilers / etc have conspired to give a significant improvement in delivered performance... for most codes.

Why You Should Care

- It's a good excuse to learn about the underlying technologies
- If you're expecting Moore's Law to make some project doable in a couple of years, you don't want to be surprised.
- A real issue for:
 - Oil, high energy physics, weather, crypto, any field in which data is growing, or in which computation demands are unbounded.

A side-note: Benchmarks

- Apologies to Mr. Twain: There are lies, damn lies, and benchmarks.
- SPECcpu, in particular, is an imperfect lens through which to view performance:
 - Vendors only publish the best score
 - Compilers improve over time, but vendors never re-run the test with new compilers on old cpus.
 - Difficult to back out “broken” benchmarks
 - Sun busted 2 benchmarks in SPEC2000fp

Rapidly Improving Technologies, Unrelated to Moore's Law

- Bandwidth
 - High speed serial copper & optics
 - Point-to-point links: PCI Express, SATA, Serial Attached SCSI (SAS), plus switching
 - Sharing of technologies:
 - DDR (memory, PCI-X 2.0, HT) / QDR (Intel FSB)
 - 1X links (IB, alternate 10gigE), 2 Gb links (FC, Myrinet)

Rapidly Improving II

- Disk Capacity: 2x every 9 months for ~ 5 years!
- Driven by physics completely different from Moore's Law
- Extremely handy for people like astronomers, whose detectors get better with Moore's Law.

Slowly Improving Technologies: Latencies

- “The one thing that doesn't scale as Moore's Law is the speed of light”
- The best example: a wide-area network
- US coast-to-coast round trip for a photon is 30 milliseconds.
- From my apartment to Virginia, 80 ms
- \implies not much improvement possible.

Slowly Improving: WAN latency II

- Fortunately, we're talking about cluster computing, not distributed or Grid or whatever
- For shorter distances, only a part of latency is caused by the speed of light
 - Room, board, disk, cache, OS
- This means they will improve some, but not very quickly

Slowly Improving: LAN Latency

- Ethernet
 - Gig switches are all store-and-forward
 - Offload engines reduce overhead, but not necessarily latency
 - The average use of Ethernet doesn't care much about latency, so it is not optimized for

Slowly Improving: LAN II

- Specialty interconnects: Myrinet, IB
 - Protocol complexity requires cpu on interface card
- As bandwidth rises, all latencies get worse
 - Example: 4X IB has to deal with skew
- Additional overhead for MPI is significant if you did manage really low latency hardware

A Short Digression into Scaling

- 2 kinds of scaling: Strong, and Weak
- Weak scaling refers to system speedup with increasing data size.
- Strong scaling refers to system speedup with constant data size.
- Everyone finds weak scaling easier to achieve than strong scaling.
- Most cluster-friendly problems have weak scaling.

Scaling Digression II

- Strong scaling involves rapidly falling latency, which is impossible to attain for long
- The true scaling of an algorithm often obscured by programming details
 - i.e. MPI weather codes are getting better at combining their messages to neighbors
- The point: Everyone seeing scaling that falls off could use lower latency interconnects, but the market isn't necessarily big enough to fund R&D

Slowly Improving: Memory Latency

- Intel system: over shared bus to North Bridge, thence to memory.
- Sharing is expensive: $\sim 2X$ latency
- Bandwidth is not your friend: RAMBUS memory has higher latency than conventional DRAM
- AMD Opteron cuts out 1 bus crossing, ties clock of North Bridge to main CPU.
- $\frac{1}{2}$ remaining latency in memory bus, $\frac{1}{2}$ in NB

Slowly Improving: TLB

- Each memory reference requires a virtual to physical address translation
- The TLB caches these translations
- x86/Athlon have a fairly fast and large TLB (good), but don't really support intermediate page sizes.
- A problem for codes with irregular data.

Slowly Improving: Disk latency

- Oh my God, it's a moving part!
- Fortunately, most clusters do large sequential I/Os, for which latency doesn't matter.
- Oracle customers are paying for improvement to disk latency anyway.

Slowly Improving: Cache

- Most scientific codes make good use of cache
 - John McCalpin has a good paper on this
 - Very unfortunate for codes that do *not* make good use of cache: irregular data
- Cache is now mostly internal to the cpu chip
 - Which means it clocks up with the CPU
 - It actually is gradually falling back, because the speed of light does matter.

Slowly Improving: OS latency

- Consider a full context switch
 - Including the time needed to heat up the caches
- Over time, this number changes slowly
- Linux is about to have a nice jump thanks to a new x86 instruction for syscalls
 - Already in use on x86_64 port
 - About to sneak into i386 glibc.

Slowly Improving: Heat Transfer

- Heat per volume is at an all-time high
- Cooling technologies become much more expensive beyond a slowly-moving point
 - Intel talk puts that ~ 80 watts/cpu
- 20 KW in a rack is hard to air-cool
 - Difficult to get enough air handlers for 10+ racks, because air can only transfer so much

Laptop chips are not the answer

- Lower performance per chip, and cooler
- Heat / chip and heat / flop are on a different curve, but with the same trend as desktop cpus
- Same situation as with blades

Desktops are not the answer

- Intel and AMD are in an arms race on the desktop.
- A desktop only has 1 cpu, so they're willing to let total power dissipation rise, if it gives a higher clock.
- The only limit is the \$ to cool a single cpu system.

Power Management is not the answer

- Clusters have pretty high utilization %, and cooling must be sized for the maximum
- People don't mind so much if their server is slowed a bit for over-temp... but they mind a lot if their lock-step cluster has random slowdowns
- Example: My Athlon machines from one of the bottom-feeder vendors slows 10% if I run an overclocker “cpuburn” program.

Gut-Wrenching Nightmares

- \$ per performance, falling rapidly
- Performance per cpu, rising rapidly
- Power consumption per performance... whoops
- A machine room full of 1U servers can be filled for a constant \$, but needs continual upgrades of cooling & power.
- A significant problem for medium to large clusters

To conclude

- Scaling issues haven't changed, for better or worse, in a while
- Facilities issues are the # 1 issue for clusters
- Clusters will continue to dominate weak scaling problems.
- Strong scaling continues to be a problem

Coda: Twain

<http://www.enterstageright.com/archive/articles/0202/0202shams.htm>

In his autobiography, Twain wrote: "Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: 'There are three kinds of lies: Lies, damned lies and statistics.'" However, no one has been able to find evidence that the British Prime Minister ever made such a remark. As noted in Ralph Keyes's *Nice Guys Finish Seventh*, Harper Collins, 1992, investigators have discovered the following comment by a member of the (British) Royal Statistical Society: "We may quote to one another with a chuckle the words of the Wise Statesman, lies, damned lies statistics..." (*Journal of the Royal Statistical Society*, 1896). The similarity between the comments is unmistakable, but the connection to Twain or Disraeli is unknown.

See also: <http://www.ling.helsinki.fi/~widenius/clemens.htm>