HP & PetaFLOPs (But first some words from your electric company)

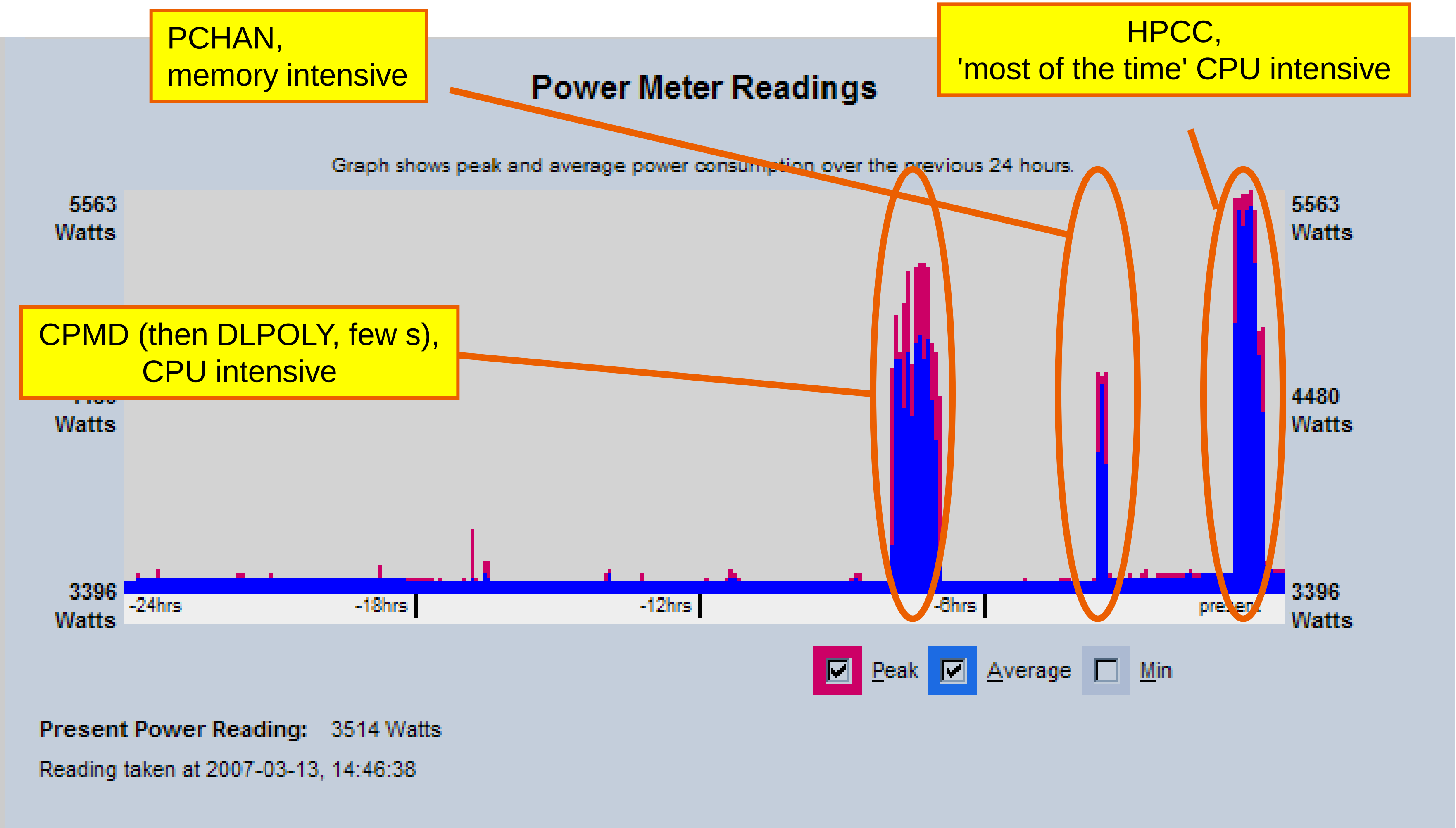
SOS 11

Challenges of Sustained Petascale

Richard Kaufmann
High Performance Computing Division, HP

Servers Have Turned Into Power-Hungry Beasts!

(And HPC workloads are the worst)



Power Meter Readings

Graph shows peak and average power consumption over the previous 24 hours.

☑ Peak ☑ Average □ Min

Present Power Reading: 3514 Watts

Reading taken at 2007-03-13, 14:46:38

PCHAN,
memory intensive

HPCC,
'most of the time' CPU intensive

CPMD (then DLPOLY, few s),
CPU intensive

Single enclosure with 64 cores @2.66Ghz

Slide courtesy of HP EMEA

Initial Purchase Price vs. 3-year TCO

- Interesting non-hypothetical question
  - Would you pay an extra US$100 for a server that had a more efficient power supply?
    - Example: ~70% efficient supplies are really, really cheap; ~90% efficient supplies aren’t
    - Assume server needs 400W, net of power supply efficiency
  - If you said yes, how much do you think you’d save over three years?
    - $0? (breaks even, but you can hold your head high knowing you did the right thing)
    - $200?
    - $400?

How about more than $600?!

Case 1: 70% Efficient Supply	Case 2: 90%
$1,500 to power the server	$1,168 to power the server
571W * 3 years	444W * 3 years
$1,700 to pay for the power infrastructure	$1,333 to pay for the power infrastructure

- US$0.10/KwHr
- US$10/W for data center costs (spread over 10 years)
  - Low end of Google spread: $10 - $22.
    http://www.eweek.com/print_article2/0,1217,a=204820,00.asp

$715 savings less $100 for the better power supply

This is one reason why (gratuitous plug: HP's) blades make sense for many HPC customers

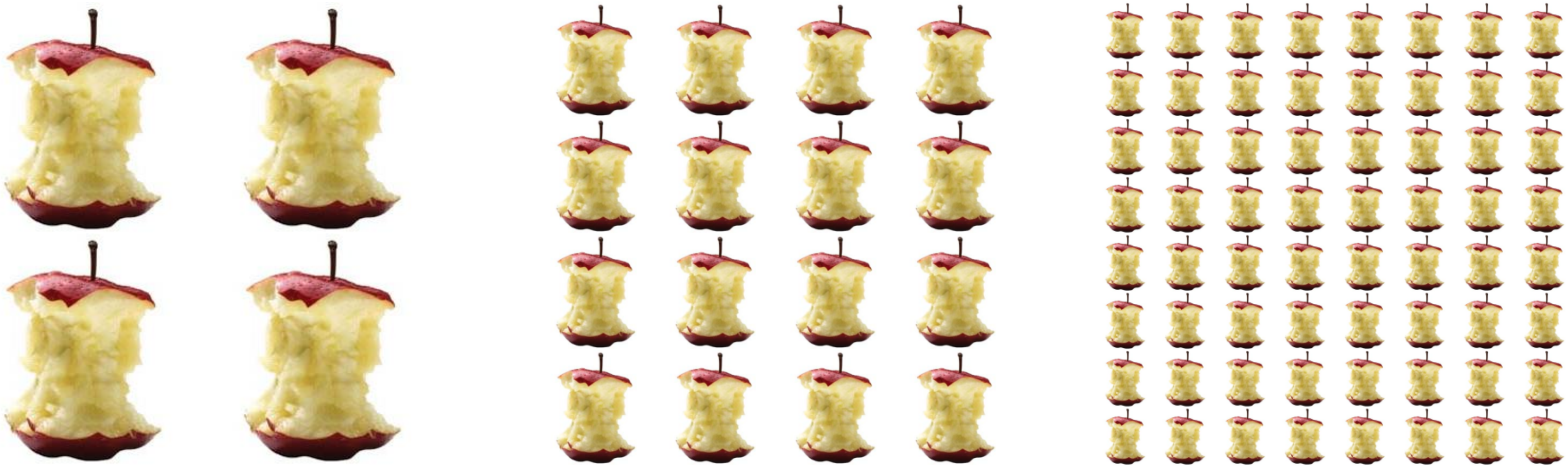
- Engineered for TCO
  - Very efficient power supplies, fans
- Redundancy without efficiency compromise
  - Power supplies run best at full load
    - Example: 3+3. Three power supplies providing the load at 100%; extra power supplies only brought on-line when required
  - Effective cooling
    - Ducted fans, "clean sheet of paper" airflow design, baffles, …
  - More connections via etch, not cables
    - 1st level network switching within enclosure
- Typical: ~25% power reduction compared to average 1U servers

TCO should be pervasive!

- You really want to be able to “pay it forward,” and select servers with (at least options for) more efficient power supplies, etc.
- Blades are built with TCO (power costs, management costs, etc.) as their top design goals

- Are you still buying systems based only on initial purchase price?

The ubiquitous foil about multi-core processors…



2007 2011 2015 etc.

Courtesy: M. McLaren

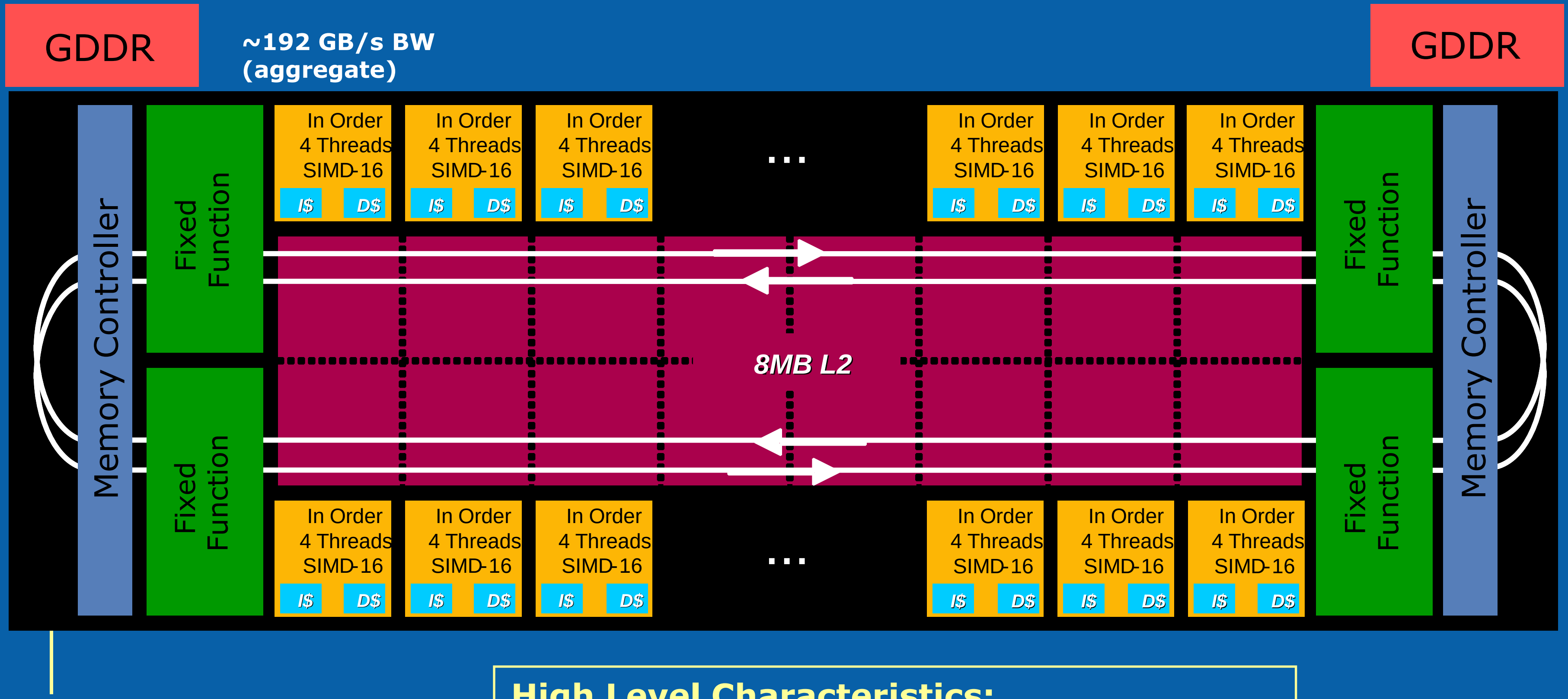
And the words to go with them…

- Cores double every eighteen months, more or less
  - No law says it has to be 2, 4, 8, …
  - Function of fab economics and user needs, not slavish devotion to powers of two
- Speculation (based on a LOT of FUD)
  - FLOP acceleration beyond that gained by increasing core count
    - Side effect of GPU wars

But How Will Those FLOPS Be Delivered?

- Per-thread performance will remain somewhat static
  - Perhaps “simplified cores” will enable core count acceleration beyond what comes with shrinkage
    - Perhaps a few “fast” cores for the stubborn threads
  - Speculation: We’re all going to get tired of “around 3GHz”
- Floating point units will get a lot more capable
  - Side effect of GPU arms race
    - TF chips need very wiiiiiiiiiiiiiiiide floating point paths
- And, of course, you’ll be up to your neck in cores!

Larrabee as a Dev Platform for Future HPC Many Core Products



~192 GB/s BW (aggregate)

~2 TB/s BW (aggregate)

High Level Characteristics:

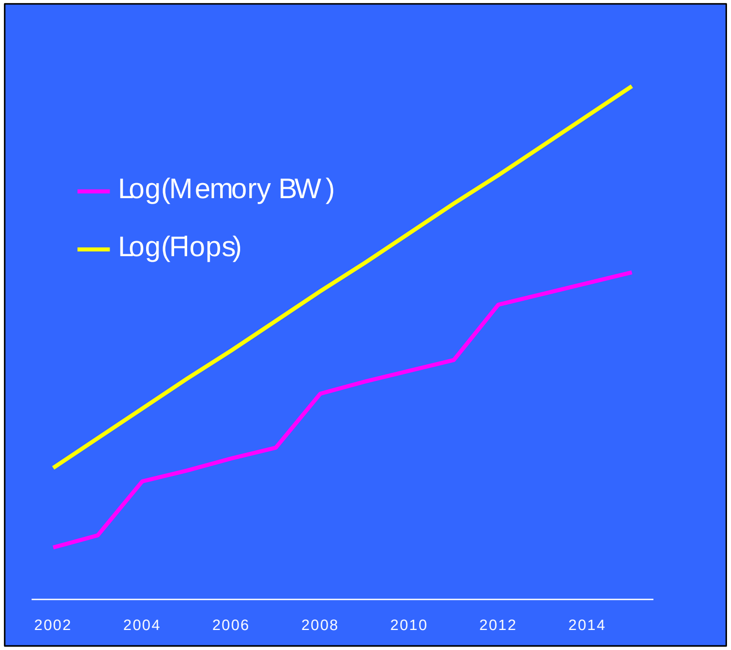
- **Many-core X86 & tightly coupled VPUs**
- **True data parallel architecture**
- **~2 TFLOP aggregate throughput**
- **Vector machine (SIMD-16)**
- **Highly threaded (128 total HW threads)**

Equal Time

- You’ll have heard all sorts of great things from AMD earlier(!)
  - Push for Torrenza: HT-based acceleration
    - HTX Slots → On Package →? On Die
  - Push for Fusion: ATI+AMD
    - GPU integration on-package? On-die?

But What Of The Memory Subsystem?

- Memory bandwidth (still) increasing incrementally over the next few years
  - Gently frequency bumps
    - DDR → DDR2 → DDR3 → etc.
  - FBD (nothing comes for free)
    - Latency, Power ++
    - Bandwidth++, pin count - -



A breakthrough is needed (optical!), but won't happen for 5+ years

Potential intermediate answers:

Additional memory channels

Mux chips (used in PA-RISC, HP Itanium)

I/O

- PCI-E → Gen2
  - Enabler for QDR IB
  - First server platforms ~end 07
- Geneseo
  - Coherent and atomic ops across PCI-E
  - Response to HT
- HT
  - HT3 Direct-attach Accelerators, NICs, …

What's Up With HP & PetaFLOPs?

- Yes, we bid with Intel, PSC and Sandia on a sustained PetaFLOP machine!
- What we can say
  - Intel ManyCore + Aggressive 3D Torus Interconnect + Next-generation blades
  - Interesting challenges
    - RAS + Packaging + Link technology
    - Programming models, tools
    - Power!
- This is a product effort
  - It'd be a heck of a "Serial 1"!
  - HP's effort addresses both high-end and ISV-led midrange market

Optical…

- Long copper IB cables are about to disappear
  - Optical E-O-E cables “real soon now” from multiple suppliers
- Tug of war:
  - Optical pushing to replace copper at shorter and shorter distances
    - Row-to-row → Rack-to-rack → Intra-Rack → Intra-Server → Server ⇔Memory → Intra-chip
  - Copper driving down cost/bit/sec
  - *Very* much like the CRT fight with flat panels
    - *1980: Prevailing opinion “Flat panels will replace CRTs in a few years”*