SOS13

Richard Kaufmann, SCI

March 2009

The Usual Reminder

- HP makes no warranties regarding the accuracy of this information. HP does not warrant or represent that it will introduce any product to which the information relates. It is presented for entertainment value only. So there.

Hypothesis 1. With users' relentless appetite for HPC we could expect systems with 100 PFlops peak soon after ~2012.

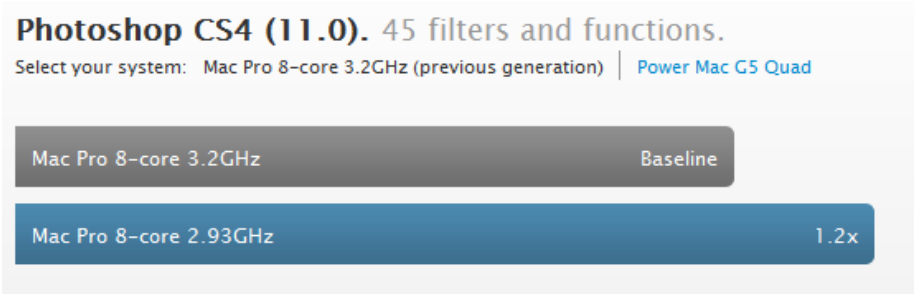
- What type of systems do you think will first achieve this?
- Will they be general-purpose?
- How much electrical power will they consume?
- What will be the standard way of programming applications for these systems?

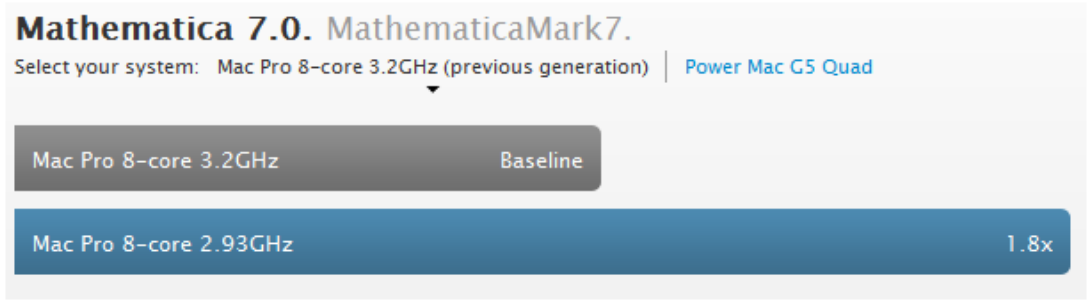
First, a mainstream calculation

- Let's say a chip is about 200GF in 2012. (Two doubles from today's 50GF parts.)
- 100PF ÷ 200GF = 500,000 parts
  - @125 watts per socket: ~60 MW
- I'm sure Intel or AMD would love to sell you the parts, but…
- Conclusion
  - VERY fragile if used as a capability machine
  - Useful?

Observation

- Mainstream CPUs are actually taking a different direction
  - Working on fundamental ratios more than peaks
  - They're prop. better at *some* things, not all things
- Nehalem (public now, courtesy of Apple…)
  - Roughly the same peak FLOPs, "2.4 X" more memory bandwidth. Result: http://www.apple.com/macpro/performance.html





Oddity: better comparison is 2.8 :: 2.26

hp invent

Programming

- It won't be called Fortran?
- Serious answer
  - HPC émigrés are perhaps more interesting than HPC itself!
    - Many web analytics apps that are mainstream today would be considered HPC apps
    - The programming model here tends to be domain specific (or tailored), e.g. map/reduce
    - *Throughput increase* is sufficient for many users
      - TCO is the key metric, and these users don't think of themselves as doing parallel computing
  - Haven't seen too much excitement for anything other than MPI…
    - Interesting opportunity: GPU-specific libraries (Grand Central, CUDA, CTM, …)
      - Doing a good job in some verticals, e.g. oil and gas
      - Within one year: essential for video editing, photo processing
    - Not arguing it's elegant for crafting TTS parallel systems beyond a node. But solutions like this remove more long-hanging fruit.

And a complaint

- HPC users typically lobby *hard* for increased memory bandwidth
  - And then brag about Linpack
- Better career move for computer architects: go for what makes the mainstream faster
  - Caveat: There is something to the idea of preparing for future generations, e.g. more cores in the same socket

Hypothesis 2. The HPC industry faces huge challenges including…

- Mounting power consumption, maintaining system availability with increasing component volumes, decreasing memory bandwidth per core, software to scale to millions of processors.
- Yet vendors seem able to re-invent solutions on a regular basis.
- What paradigm shifts, if any, do you see occurring by 2012?
- Where might you look for new partners?

Harder To Find Efficiencies

	5 yrs. ago	Today
PUE	2, 3, Higher	1.5 Good 1.3 Great
UPS Efficiency (Part of PUE)	94%	99%
Power Supply Efficiency	75%	92%
Fan Power per 2s Node	60+ W	5-10 W

Are we ignoring the elephant?

- Trying to get a bit more efficiency in a power supply is hard
- Why don’t we move the monster datacenters to the power plants?
  - Overcoming distribution losses is technically easy, but co-tenancy is organizationally difficult

HP POD:
Performance-Optimized Datacenter

HP POD Complements Brick-and-Mortar

PUE of ~1.6

PUE of <1.25

150 - 200 W/sq.ft.

1800+ W/sq.ft.

~2 years to design & build

Six weeks to ship

Maximum Security *IT Flexibility* *Max Redundancy* *Energy Efficiency* *Geographic Flexibility* *Power density* *Speed of deployment*

Brick-and-mortar — *Container*

hp invent

HP POD Key Features

Class-leading Industry-standard Flexibility

22 x 50U, 19” full-depth industry-standard racks support HP, Dell, IBM, Sun, Cisco, etc.

Best-in-class Density

Support for 3,520 compute nodes, 12,000 LFF drives, or any combination

Built-in Redundancy

Power and cooling redundancy, including separate power feeds to the racks

Ships from factory in 6 weeks, deployed WW

Pre-integrated, configured and tested before shipment; shipped in six weeks from order.

Energy Effectiveness

PUE ratio <1.25 (1.07 excluding chiller)

Infrastructure Services Portfolio

Full lifecycle support services combining technology and facilities expertise
From site preparation to ongoing POD maintenance and support

Interior view

Serviceable high efficiency heat exchangers (HEX) from HP MCS

Serviceable high efficiency, variable speed blowers from HP MCS

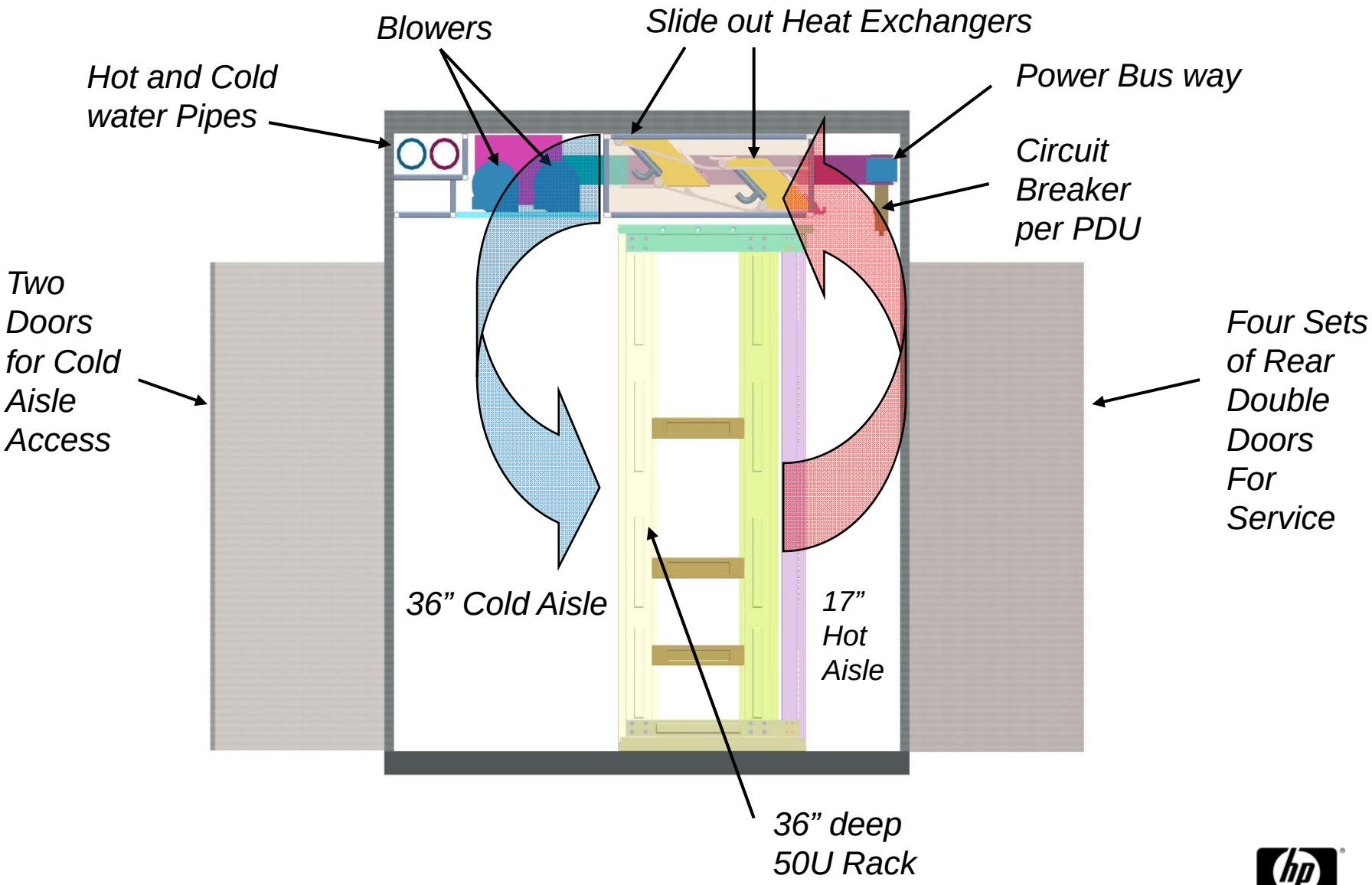
Standard 50U racks

Facilities management on exterior of cold aisle

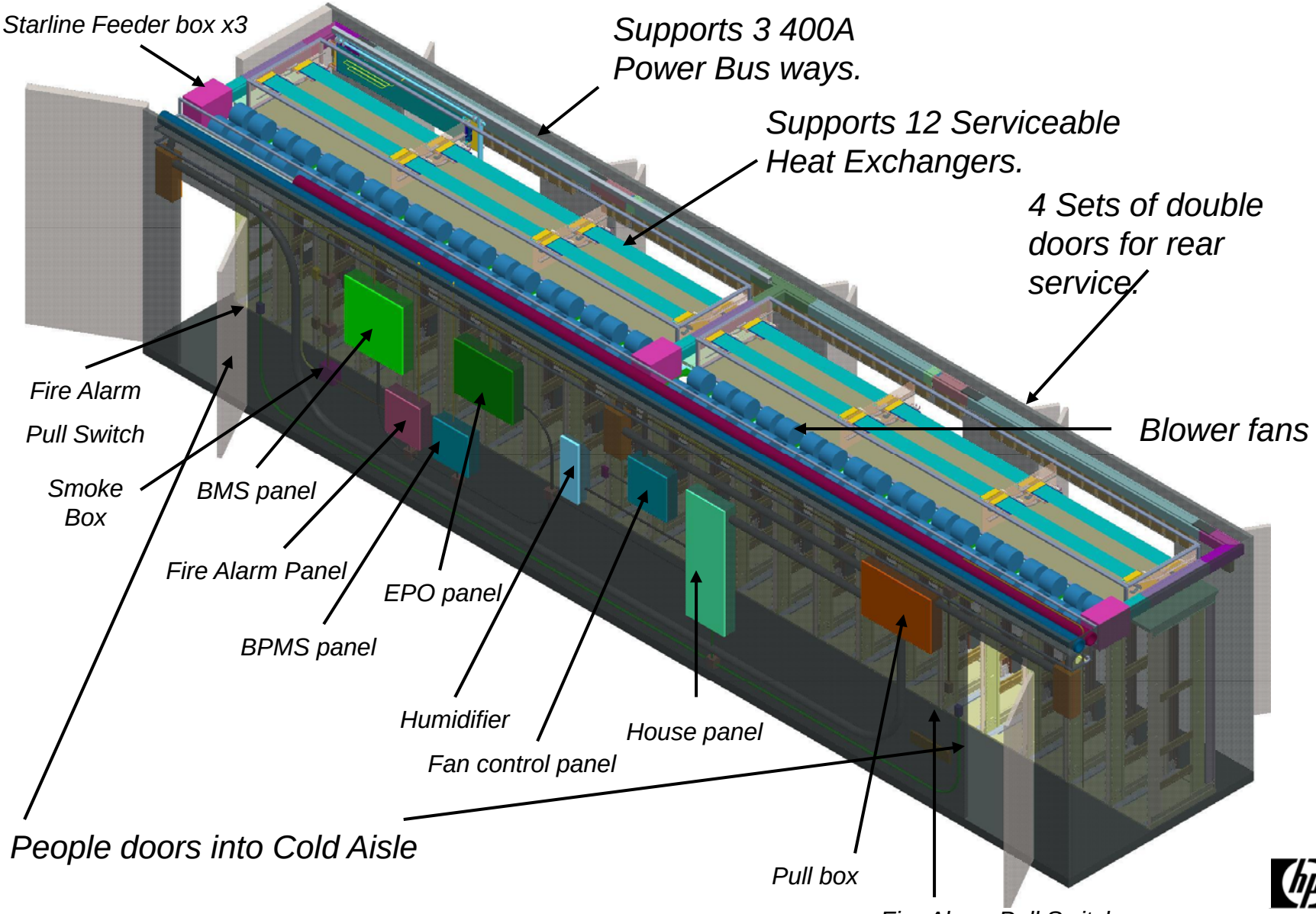
Hot aisle with rear access through doors in the container

36” cold aisle can run at >90F

End Detail View



40 Foot Container Layout



Thinking Beyond HPC…

- Lots of commonality between cloud and HPC computing
  - Node types, scale-out software models
    - Cloud: lots of innovation, fresh ideas, …
  - Drives HUGE volume

Constrained Differentiation in the Mainstream

- Customers reward designs that:
  - Minimize TCO via ease of management
  - Integrate technology in a way that increases RAS
  - Less power, more density
  - Are delivered TTM with the base technology
- Innovation constrained (but highly efficient) due to the horizontal nature of the business
- Historical perspective: mainstream technology has steamrollered “outliers” with grim efficiency
  - Being different means always being one mistake away from oblivion

Use cases: Workloads

- Search
- Web Tier
- Database Servers
- Traditional App Servers
- Cache Nodes
- “HPC” Compute Nodes
  - e.g. 2*QC, 2GB/core, 95W SKUs, one PCI-E 8X slot for IB, checkpoint/restore model places medium-high value on node reliability, density useful because of large clusters (& sometimes cable lengths)

Requirements Drive Server Designs

- Socket count
  - 2s has a lot of user acceptance, regardless of core counts
  - “The herd” likes to move together
- Memory / Socket
  - Going up with core count for some applications
- Slots
  - “Glued down” NICs may reduce the need for slots for many users and many uses
    - Accelerators a notable exception
- Drives / server, type, RAID or JBOD, internal or external,

Use Cases: Datacenters

- Datacenters with10KW/rack: current “co-lo” sweet spot
  - Many co-los still at 5KW
  - Some newer centers at ~15KW
  - PUEs ranging from 1.3 – 3.0 (PUE = total watts / watts doing stuff)
- HPC customers often aim much higher
  - Fully populated blade racks: 32KW+
- 2011
  - Scenario 1: 50KW+ in a datacenter. Or maybe not!
  - Scenario 2a: Extremely dense containers. High power density, water-cooled. (PUE of 1.2 or better).
  - Scenario 2b: Containers cooled with outside air: less dense, even better PUE (ref: Intel experiment)

PUE Definition: www.thegreengrid.org

Use Cases

- Horses for courses
  - Redundant power supplies/fans, hot swap disks, CPUs, memories, … can make sense for some applications
  - Other applications want the cheapest node possible, and are willing to accept node failures *and even wrong answers* up to a point.
  - Trend: reliability implemented in software, with less need for reliable hardware
    - Example: cloud file services use replication instead of RAID
    - Hard to do
    - Impractical for some applications
    - Only goes so far…

Hypothesis 3. The current world economy is drastically different to any ever seen in the lifetime of the HPC industry.

- Credit may effectively disappear, funds may become more centrally-controlled, hyperinflation may arise from capital injections, and markets shrink.
- Yet HPC users benefit from healthy competition sustained by the current market size.

Server Mfgrs Were Already Following The Willie Sutton Model

- (Why do you rob banks, etc.)
- How much profit is in the top 10 scientific computing systems (as measured by any metric you'd like)?
- Trends (personal projection)
  - More rapid transition to a horizontal ecosystem
    - Some folks make chips, some make servers, some make scale-out software stacks, etc.
    - More efficient, but aggressively punishes outliers
  - Market consolidation
    - When a company's market cap is smaller than the cost of *one* delivered system…

What do you think is your company's best strategy for survival?

How do you think HPC customers can realistically help you?

- (You) continue buying the best systems for your applications.
  - The more the better!
- (We) focus on "stretching" mainstream technologies.
  - E.g. High performance interconnects
- (We) invest in key base technologies that can break the memory wall, etc.
  - Si Photonics, Memristor, …