The Six-Million Processor System

Wu FENG

feng@lanl.gov

Los Alamos National Laboratory

NNSA

Los Alamos
NATIONAL LABORATORY

Abstract

... ASC Kaleidoscope, Large-Scale System ... A system "barely" alive ... Gentlemen, we can rebuild it ... we have the technology. We have the capability to make the worldÍs first six-million processor system. ASC Kaleidoscope will be that system. Better than it was before ... stronger, faster ..."

This tongue-in-cheek introduction, courtesy of "The Six-Million Dollar Man" TV show, purports that the current approach to building large-scale systems for scientific computing is flawed and on life support but that we have the technology and ingenuity to build something much better.

Mission of This Panel

How would you build the world's first six-million processor system?

Panelists

- C. Gordon Bell, Microsoft Research
- Allan Benner, IBM
- Carl Christensen, University of Oxford
- Satoshi Matsuoka, Tokyo Institute of Technology
- James Taft, NASA Ames Research Center
- Srinidhi Varadarajan, Virginia Tech

Ground Rules for Panelists

- Each panelist gets SEVEN minutes to present his position (or solution).
- Panel moderator will provide "one-minute-left" signal.
- During transitions between panelists, one question from the audience will be fielded.
- The panel concludes with 30-40 minutes of open discussion and questions amongst the panelists as well as from the audience.

Technical Issues to Consider

- What will be *the* most daunting challenge in building such a system?
- What are the challenges in building such a system relative to
  - Hardware and architecture
  - System software
  - Run-time system
  - Programming model
  - Administration and maintenance
  - Infrastructure (i.e., how do we house such a system?)

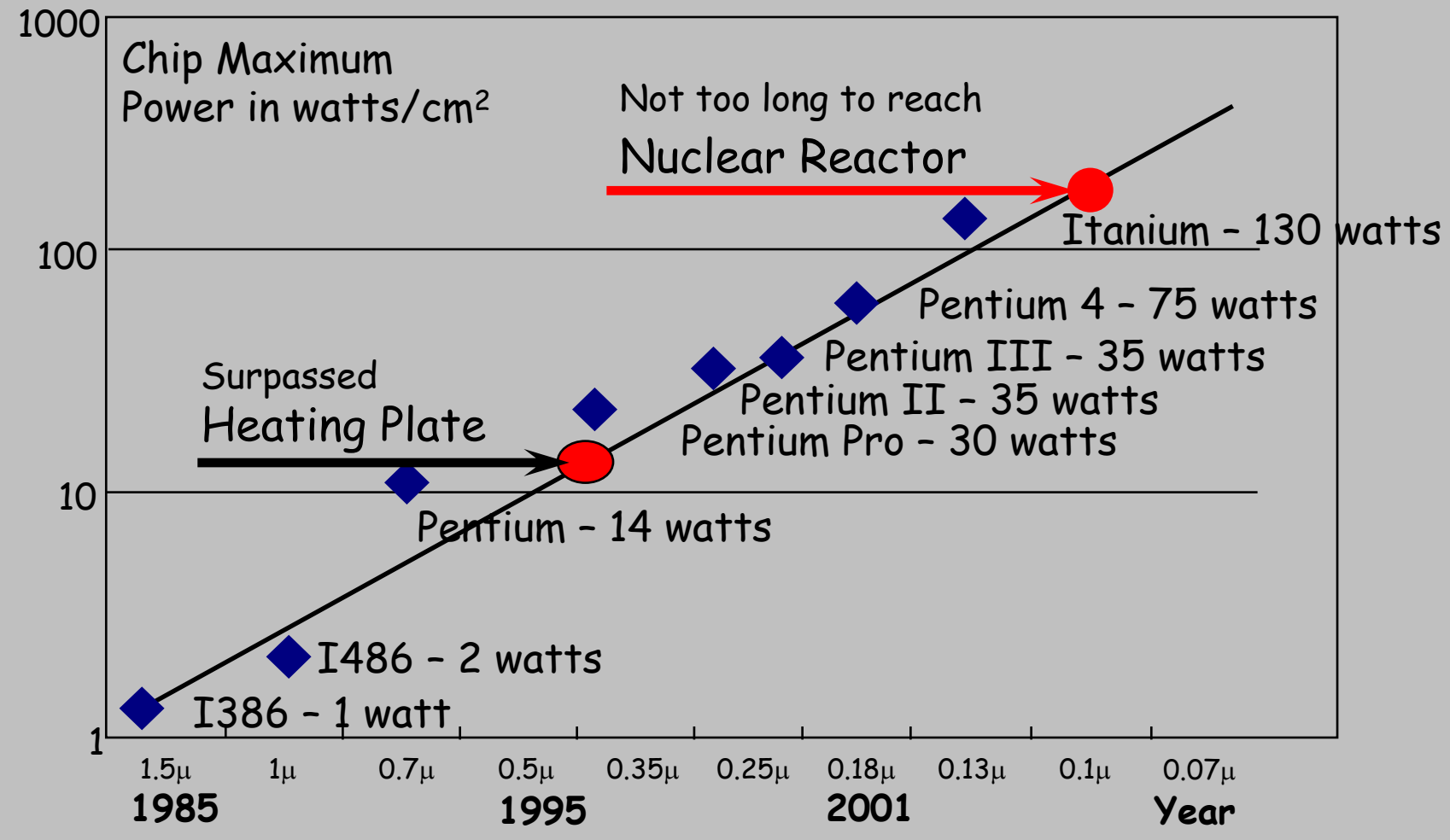
Philosophical Issues to Consider

- What *is* a processor?
  - General-purpose CPU, FPGA, SIMD processor (*a la* Cell with its synergistic processors)?
- Performance
  - Is it really just about speed?
  - What about other metrics?
    - Efficiency, reliability, availability, and scalability
    - Fault tolerance
    - Ease of administration and maintenance
    - Programmability
    - Power consumption
    - Acquisition cost versus total cost of ownership

My Two Cents?!

- Power, Power, Power!
  - Moore's Law for Power Consumption
  - Operational Costs: Power & Cooling
  - The Effect on Reliability & Availability

Moore's Law for Power



Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta

Wu FENG
feng@lanl.gov

Operational Costs of a 6M-Processor Supercomputer

- Power
  - Example: ASC White
    - 2 MW to power, ~2 MW to cool.
  - $0.13/kWh
    - $520/hour → $375K/month → $4.5M/year
- Crude Extrapolation of ASCI White
  - Assumption: Processor = General-Purpose CPU
  - 8192 CPUs → 6,000,000 CPUs
  - Power: 2930 MW = 2.93 GW (i.e., > Hoover Dam)
  - $380,859/hour → $274M/month → $3.3B/year

Reliability & Availability of Leading-Edge Supercomputers

Systems	CPUs	Reliability & Availability
ASCI Q	8,192	**MTBI: 6.5 hrs.** 114 unplanned outages/month. - HW outage sources: storage, CPU, memory.
ASCI White	8,192	**MTBF: 5 hrs. (2001) and 40 hrs. (2003).** - HW outage sources: storage, CPU, 3rd-party HW.
NERSC Seaborg	6,656	**MTBI: 14 days. MTTR: 3.3 hrs.** - SW is the main outage source. **Availability: 98.74%.**
PSC Lemieux	3,016	**MTBI: 9.7 hrs.** **Availability: 98.33%.**
Google	~15,000	**20 reboots/day; 2-3% machines replaced/year.** - HW outage sources: storage, memory. **Availability: ~100%.**

MTBI: mean time between interrupts; MTBF: mean time between failures; MTTR: mean time to restore

Source: Daniel A. Reed, UNC

Observations

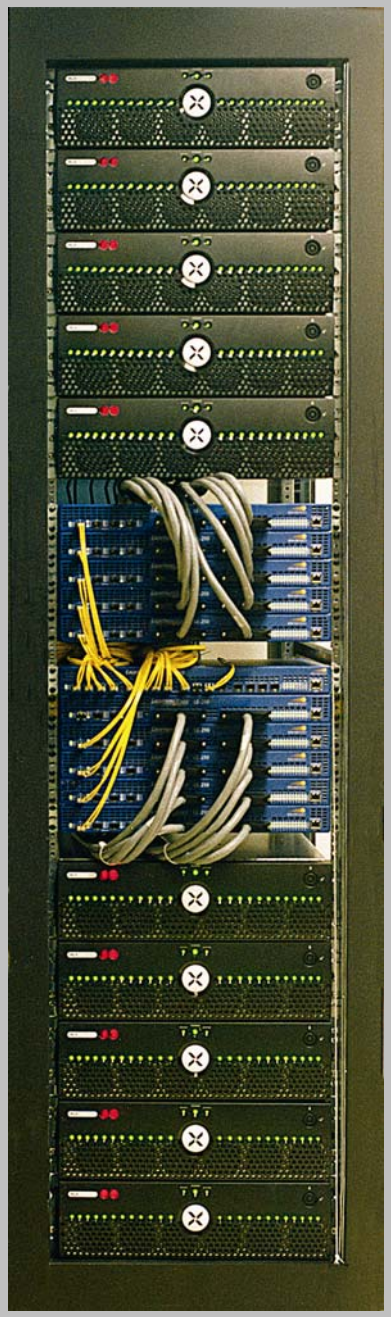
- High power density α high temperature α low reliability
- Arrhenius' Equation*
  (circa 1890s in chemistry → circa 1980s in computer & defense industries)
  - As temperature increases by 10°C …
    - The failure rate of a system doubles.
  - Twenty years of unpublished empirical data (as well as our own informal empirical data).

  * The time to failure is a function of $e^{-Ea/kT}$ where Ea = activation energy of the failure mechanism being accelerated, k = Boltzmann's constant, and T = absolute temperature
- Bladed Beowulf cluster in 85°F warehouse: Wrong answer. Move cluster to 65°F machine room: Right answer.

Commentary

- Eric Schmidt, CEO of Google
  - *The New York Times*, September 2002

    What matters most to Google "is not speed but power - low power, because data centers can consume as much electricity as a city."
  - That is, though speed is important, power consumption (and hence, reliability and availability) are more important.



GREEN DESTINY – 2003 R&D 100 AWARD

Los Alamos National Laboratory

ENERGYGUIDE

Model: Green Destiny
with High-Performance
Code-Morphing Software
Speed: 240 Gflops

High Efficiency Supercomputer
with 6 sq. ft. footprint
Memory: up to 270 Gbytes
Storage: up to 38.4 Tbytes

Compare the Energy Use of this Computer with Others Before You Buy.

This Model Uses
5.2 kWh/hr

Energy use (kWh/hr) range of all similar models

Uses Least Energy	Uses Most Energy
5.2	5000

kWh/hr (kilowatt-hours per hour) is a measure of energy (electricity) use. Your utility company uses it to compute your bill. Only models with similar performance and the above features are used in this scale.

Computers using more energy cost more to operate. This model's estimated hourly operating cost is:

44¢

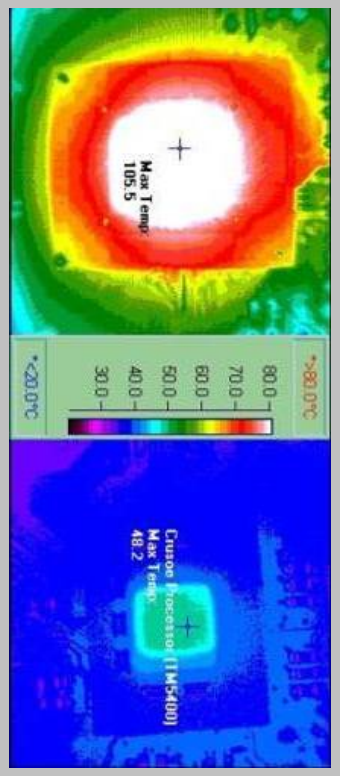
Based on a 1998 U.S. Government national average cost of 8.42¢ per kWh for electricity. Your actual operating cost will vary depending on your local utility rates and your use of the product.

Make no mistake, this is not a real label – but the info sure is real!

SUPERCOMPUTING in SMALL SPACES · http://sss.lanl.gov

Supercomputing for the rest of us!

http://sss.lanl.gov



EnergyGuide for Supercomputers?

Six Million Processors

18 November 2005

Gordon Bell

Bay Area Research

Background

- 1986 CISE Director. Parallelism challenge to CS community
- Bell Prize was initiated: 10@$2.5K; 8@$5K
- 3 approaches: vectors, MPI, u-tasking for SMP
- GB goal: 100x by 1995; *implied 1K or so in 2 decades.*
- Others described a goal of 10^6 in 15 years.

- Result: little or no creativity and help from CS community!
- 2005—wintel panic: what will we do with these new chips?
  - Multi-cores: shared cache; non-shared cache
  - Multi-threading… so far, poor designs
- Servers do ok at 32+ way SMPs.
  - Supercomputing community… no big deal
  - CS community: Let's build a new language… or whatever

Szalay, Gray, Bell on Petaflops

- Advice to NSF… and for our 6 million processors…
- Distribute them in 3 tiers… 2 million each
  - 1/3 to a few central systems
  - 1/3 or about 10x to departmental/group level systems
  - 1/3 or about 100x to individuals
- Balance the systems: 1/3 for proc, storage, & nets

Four, 2 million processor supers

- Allocate 2Mega-P to 4 center
- 512K-P centers…
- O(2^5) processors per node @ $2-3K
- 16K nodes … $32-50M = 1/3 the cost
- $100-150M systems
- Assume 4-10 GF/node… peak of 2-5 PF

The dilemma: exploiting many processors

- Wintel: Business model have been based on utilizing more capacity to provide more functionality.
  - Multi-cores: shared cache; non-shared cache
  - Multi-threading… so far, poor designs
- We see 100 processes --all interlocked on a PC
- Servers do ok as 32+ way SMPs.
  - Supercomputing community… no big deal
  - CS community: Let's build a new language… or whatever
- Matlab, databases parallelize. Program in Excel.
- I'm doing my part with Memex, a database. But do we need more processors versus more disk i/p?
- Two processors may be the norm for a decade for all but the largest, central systems e.g. MSN

The Six-Million Processor System

Alan Benner
Systems & Technology Group
Server/Network Technology & Architecture
845-433-7561 -- bennera@us.ibm.com

Nov. 18 2005

Let's Start with the most important basics:

- CPU Power & Performance
- CPU Architecture
- Memory Hierarchy
- Interconnect Technology
- Interconnect Topology
- Storage Hierarchy
- Storage Technology
- Coherence Boundaries
- Latency Tolerance
- Operating System Model
- Programming Model
- First-Level Packaging
- System-level Packaging
- Density/Interconnect Tradeoff
- Power Delivery – chip level
- Power Delivery - board level
- Cooling – Chip level
- Cooling – Board & Rack-level
- Cooling - Room-level
- Reliability
- Usability
- Administration
- Cost
- ….
- …

Cost

How much should we expect to invest in such a machine?

We have to include all the costs

- Processors – 6M
- Memory – 0.5-5GB / μP
- Disk – 10-100GB / μP
- Interconnect – 1-10 Gb/s / μP
- Storage network – 0.1-1 Gb/s / μP
- Global network I/F - 0.01-0.1 Gb/s / μP
- Cards – xx / μP
- Cables -- xx / μP
- Racks -- xx / μP
- Data Center (raised floor, RACs,…) -- xx / μP
- Building to hold the Data Center– xx / μP
- Ongoing Power Delivery (4-8 years) - xx / μP
- Ongoing System administration - xx / μP
- Ongoing Storage Administration - xx / μP
- Ongoing Access Management (allocation of computing resources to users, security, ,…) - xx / μP
- ……

Fully loaded, over the lifetime of the system, you have to assume a reasonable amount of money per processor:

~$1K / CPU

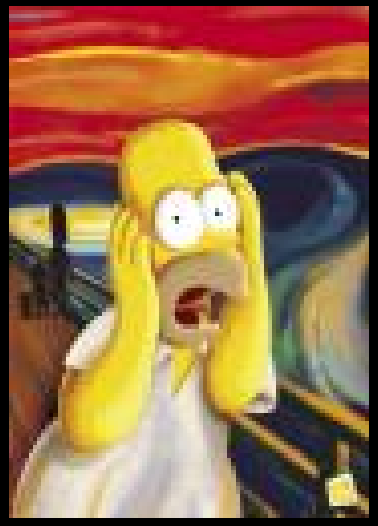
(maybe rounding a bit)

The $6 Billion Dollar System - $6B

Alan Benner
Systems & Technology Group
Server/Network Technology & Architecture
845-433-7561 -- bennera@us.ibm.com

Nov. 18 2005

But before you say:



"6 Billion Dollars!!! …Aaiiaghha!!"

Or say:

"This guy definitely wants to sell me something…."

Consider

- This would be a resource of national – or, actually international – scope, in the 21st century, with benefits going to all citizens.

- It's really a matter of setting priorities, at a national & international level

….so, where are our priorities?

One pair of alternatives

$6B Investment in Computing & Communications

- Potential Benefits
  - Finding cures for all diseases
  - Preventing all environmental disasters
  - Free energy forever
  - Movies and entertainment now unimaginable
  - Immediate access to all knowledge
  - Understanding how universes start and end
  - Understanding consciousness
  - …
  - …

5 more B-2 Bombers @ $1.2B ea.

- Potential Benefits
  - More ability to deliver large amounts of both conventional and nuclear munitions anywhere on the planet

Other ways to consider funding a $6B system

- There's a "$23.5-billion market for chocolate candy, hard candy, soft candy, mints and gum, covering both the mass-market and gourmet levels
  - http://www.packagedfacts.com/pub/143461.html
- "[Datamonitor's] report … examines trends in the $78 billion U.S. beer, cider, and Flavored Alcoholic Beverages (FABs) market
  - http://www.realbeer.com/news/articles/news-002550.php
- In Fiscal Year 2005, the U. S. Government spent $352 Billion … on interest payments to the holders of the National Debt.
  - http://www.federalbudget.com/

Allocate ¼ of what we currently spend on candy and gum

Allocate 1/10th of what we currently spend on wine coolers and beer

Allocate 1/50th of what we currently spend on interest payments for prior debt

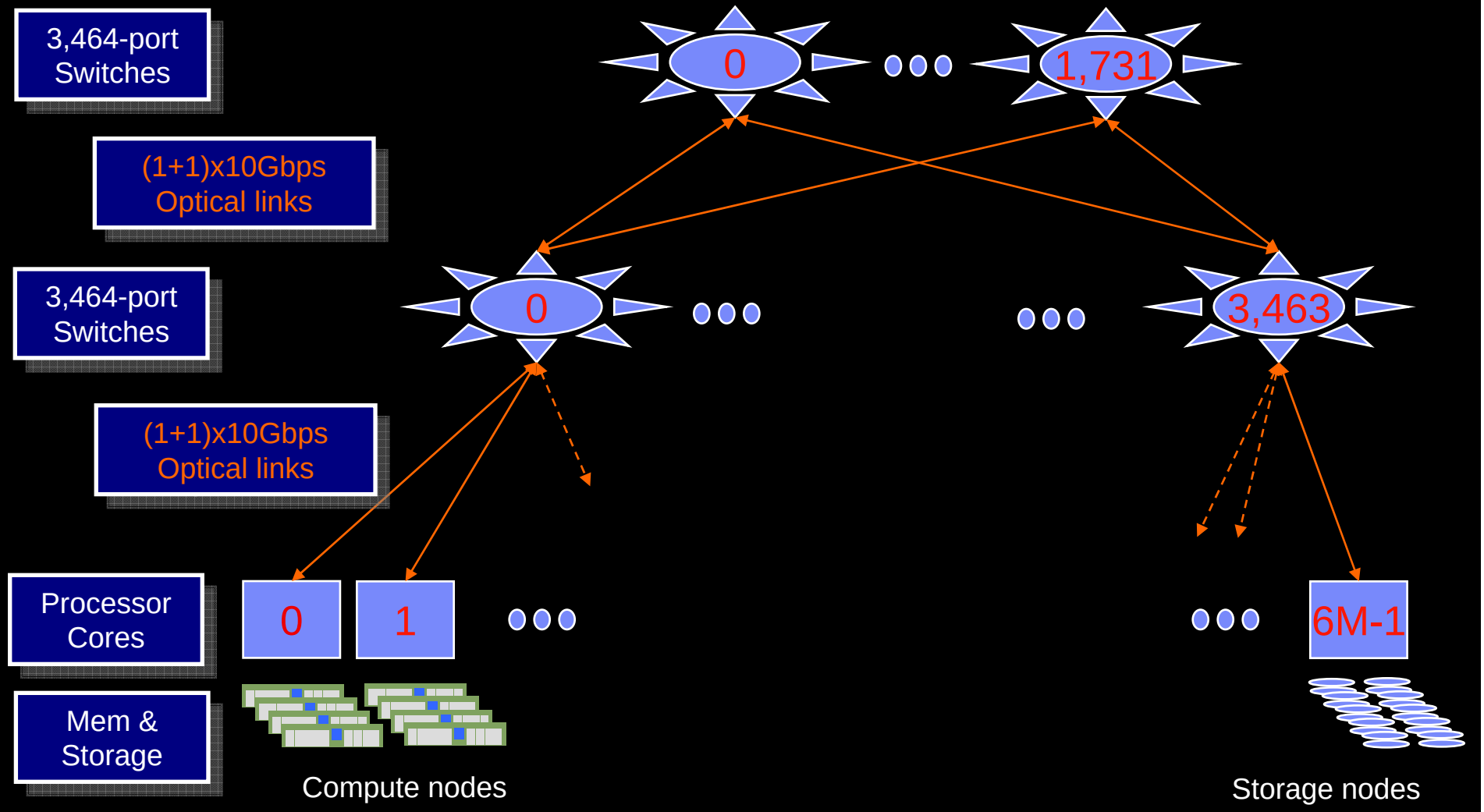
Net: Given appropriate justification for national re-prioritization of resources, we do have the money to build a $6B system.

How to build it, Take 1: First, spend some money

- $2B worth of DRAM
  - In 2010-2011, this will be ~24 PB, or 4GB/processor
  - Equivalent to 24 million DIMMs
- $1.5B worth of optics for 100 meter links
  - e.g., (12+12)-fiber XCVRs, VCSEL/MMF, 10Gbps / fiber
  - At $80/fiber, this would allow (12+12)M fibers – 2 pairs/CPU
- $1.5B of storage (hard disks & RAM disk cache)
  - Equiv to ~2 disks, $125/disk, per processor core: ~4 PB total
- $1B for Switches, Boards, Power supplies, DC/DC Converters, Racks, Cooling, Data center, Building for the data center, power plant to power the system,…, and 6M processor cores

How to build it, Take 1: The Powerpoint way

…Easy, eh?



How to build it, Take 1: Now what about space?

- Compute racks: Limited by DRAM packing density in racks
  - Blue Gene/L : 1,024 DIMMs/rack – very tightly packed memory
  - Assume 4x better memory density packing inside racks (bigger DIMMs, more tightly spaced), then 24M DIMMs will need ~3,000 Compute Racks
- Switch Racks: Limited by fiber connector edge density
  - Current switches: 1152 fiber pair in ~1/2 rack (288-port IB-4x, 4+4 fibers/pair)
  - Assume 8x better fiber connector density, a ~4K-port switch, with (1+1) fibers/pair, would take ¼ rack, so 5,000 switches need ~1,250 Switch Racks
- Storage Racks: Limited by volume of disks
  - We'll need ~1,500 racks for disk drives, too.
- Total: ~5,000-6,000 Racks, at ~2.5 sq. meters/rack = 15,000 sq. meter, or 100 meters X 150 meters
  - So, we'll put it in a building the size of a normal everyday football stadium

How to build it, Take 1: Now what about power?

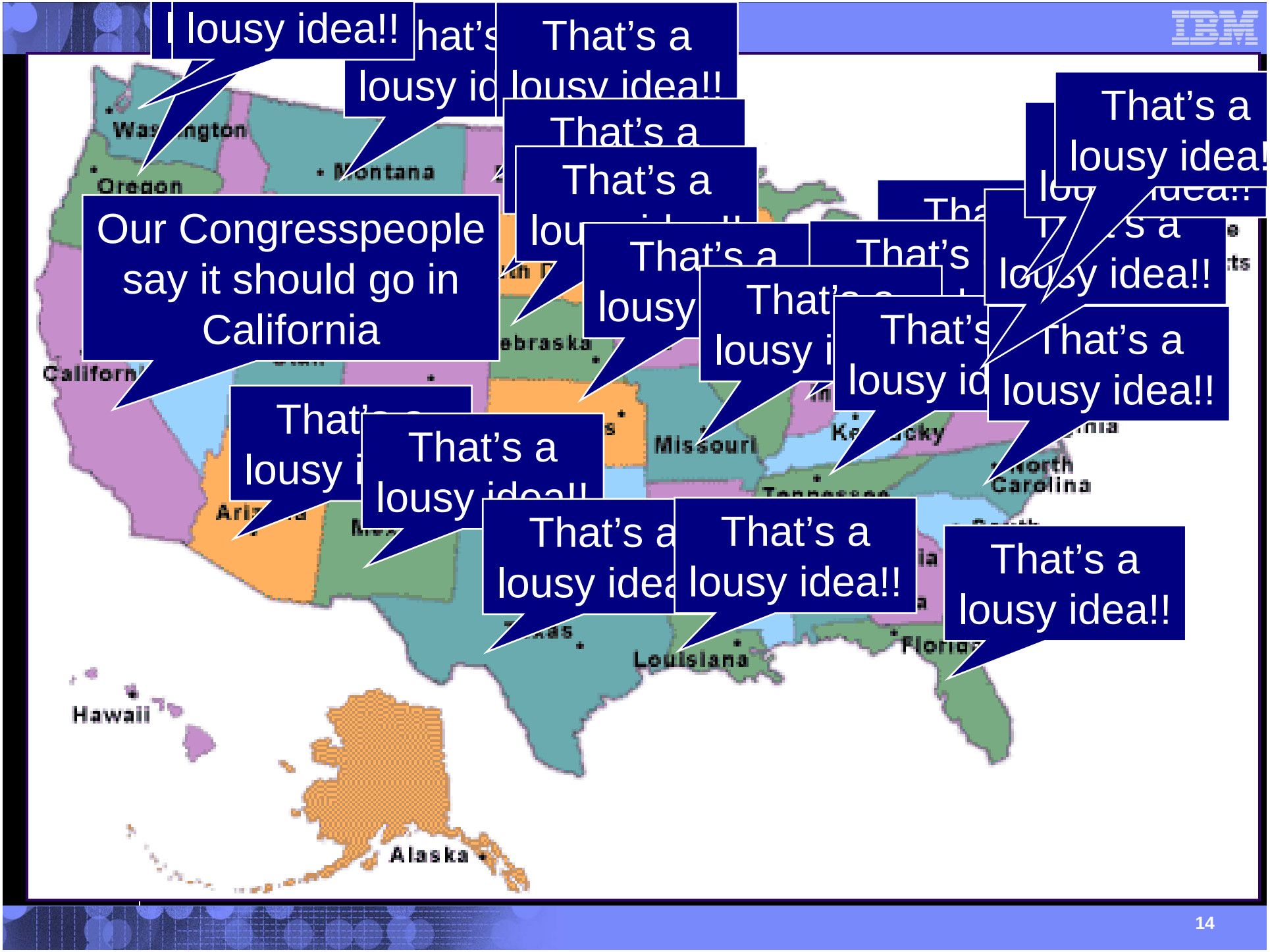
- Power per CPU:
  - CPU power: Power-efficient version ~10W
  - DRAM power: 4 1GB DIMMs, at ~20W ea: ~80W
  - Network: (10+10) Gb/s @ 0.2W/Gbps (optics+switch) ~2W
  - Storage: 2 disk drives, 2-3 W/drive ~5W

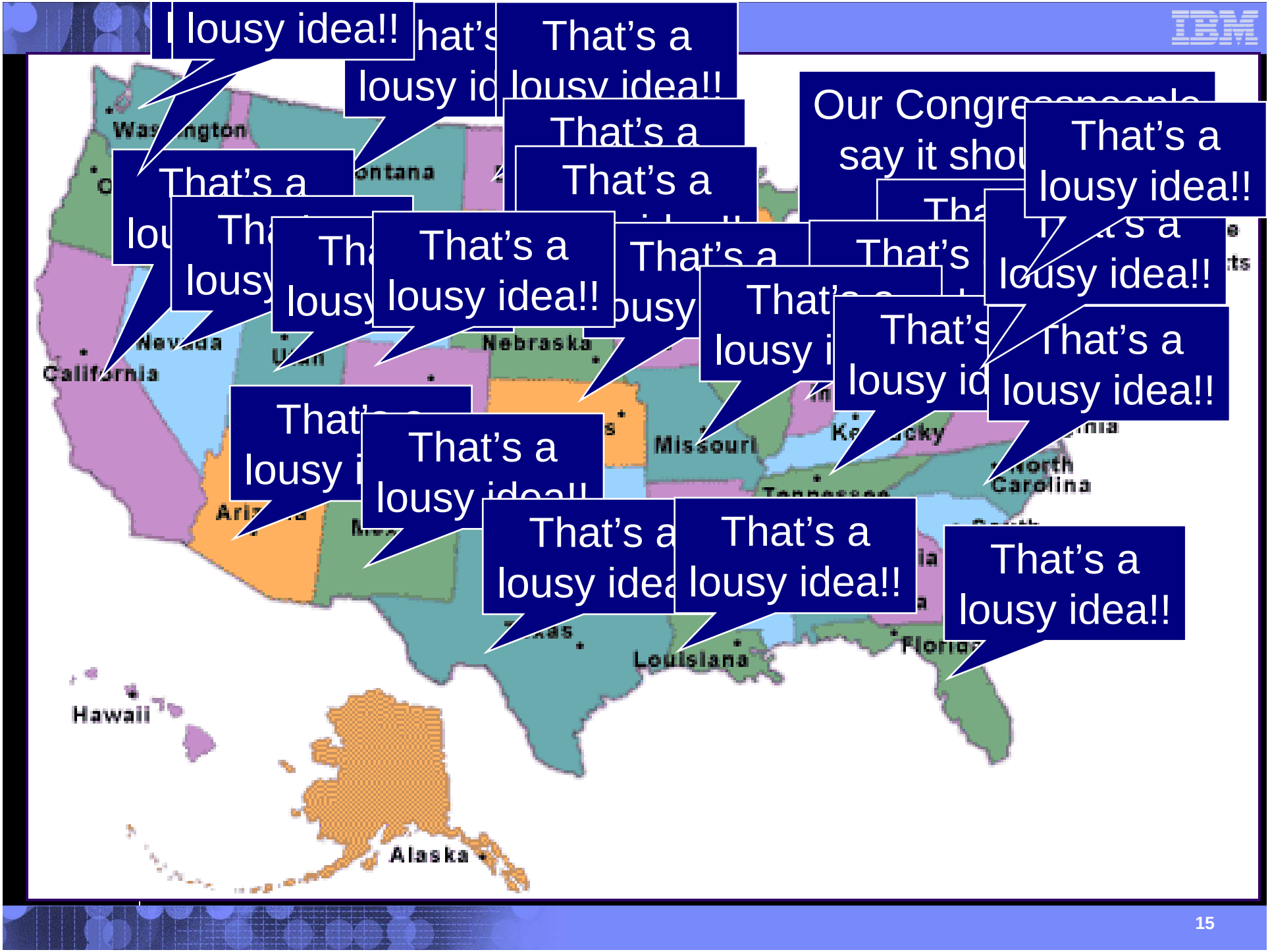
  Total: ~100W
- Aggregate system power: 600 MW
- This is only the power of a small city:
  - In 2003, New York City's … peak electricity demand was 11,020 megawatts. (http://www.nyc.gov/html/om/pdf/energy_task_force.pdf)
  - This would be only 1/20 of the power of NYC
- Again – This would be a national & international resource

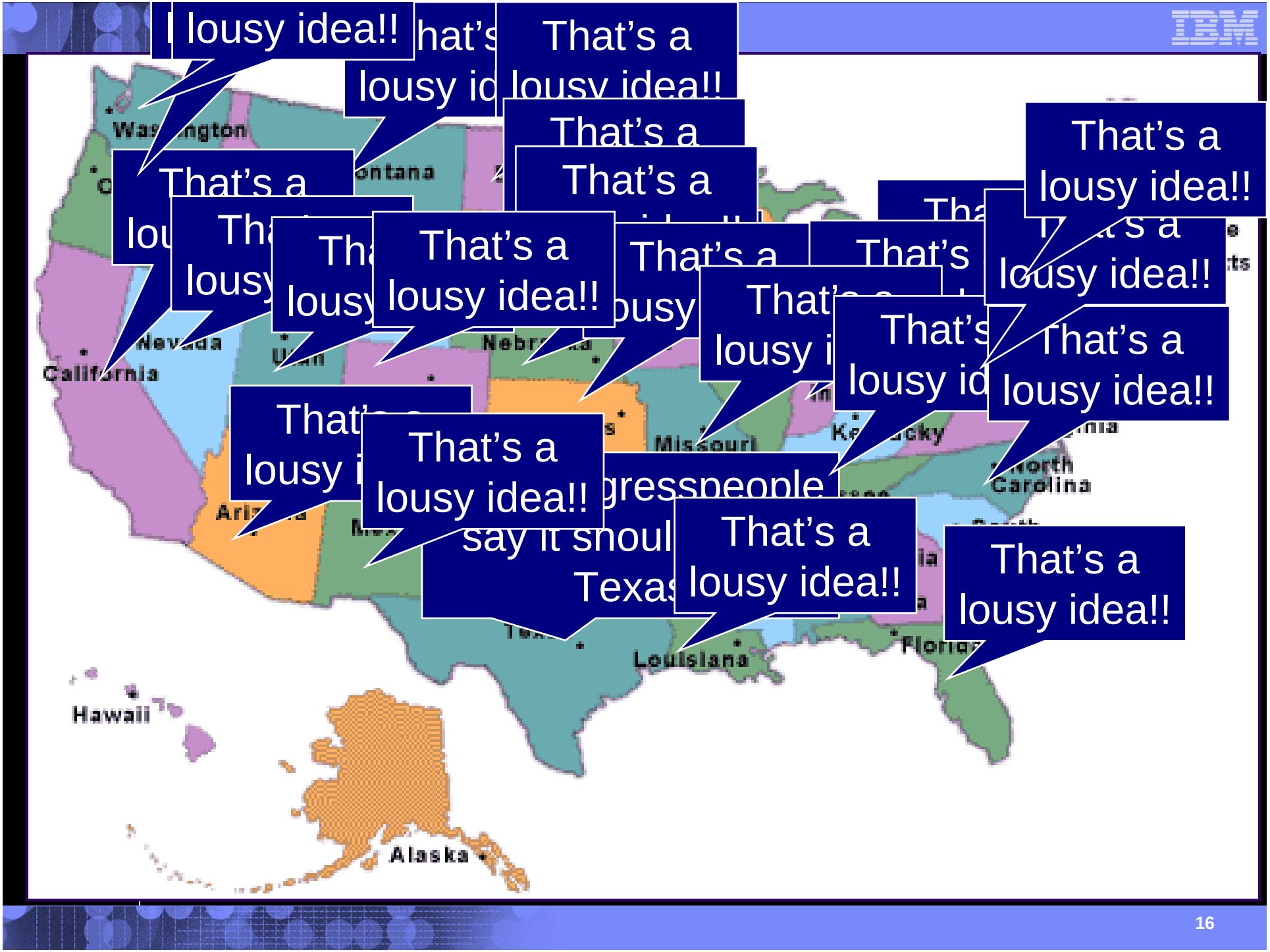
Now for the really hard problem:

- CPU Power & Performance
- CPU Architecture
- Memory Hierarchy
- Interconnect Technology
- Interconnect Topology
- Storage Hierarchy
- Storage Technology
- Coherence Boundaries
- Latency Tolerance
- Operating System Model
- Programming Model
- First-Level Packaging
- System-level Packaging
- Density/Interconnect Tradeoff
- Power Delivery – chip level
- Power Delivery - board level
- Cooling – Chip level
- Cooling – Board & Rack-level
- Cooling - Room-level
- Reliability
- Usability
- Administration
- Cost
- ….
- …

Politics!!


lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
Our Congresspeople say it should go in California
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
Oregon
Montana
Missouri
Louisiana
Hawaii
Alaska


lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
That's a lousy idea!!
Our Congre say it shou
Nebraska
Missouri
Carolina
Louisiana
Hawaii
Alaska



lousy idea!!

hat's
lousy id

That's a
lousy idea!!

That's a

That's a

That's a
lousy idea!!

That's a
lou

Tha
lousy

Tha
lousy

That's a
lousy idea!!

That's a
lousy

That's

That's a
lousy idea!!

That's a
lousy i

That's
lousy id

That's a
lousy idea!!

That's a
lousy i

That's a
lousy idea!!

gresspeople
say it shoul
Texas

That's a
lousy idea!!

That's a
lousy idea!!

The net conclusion:

This machine cannot be built

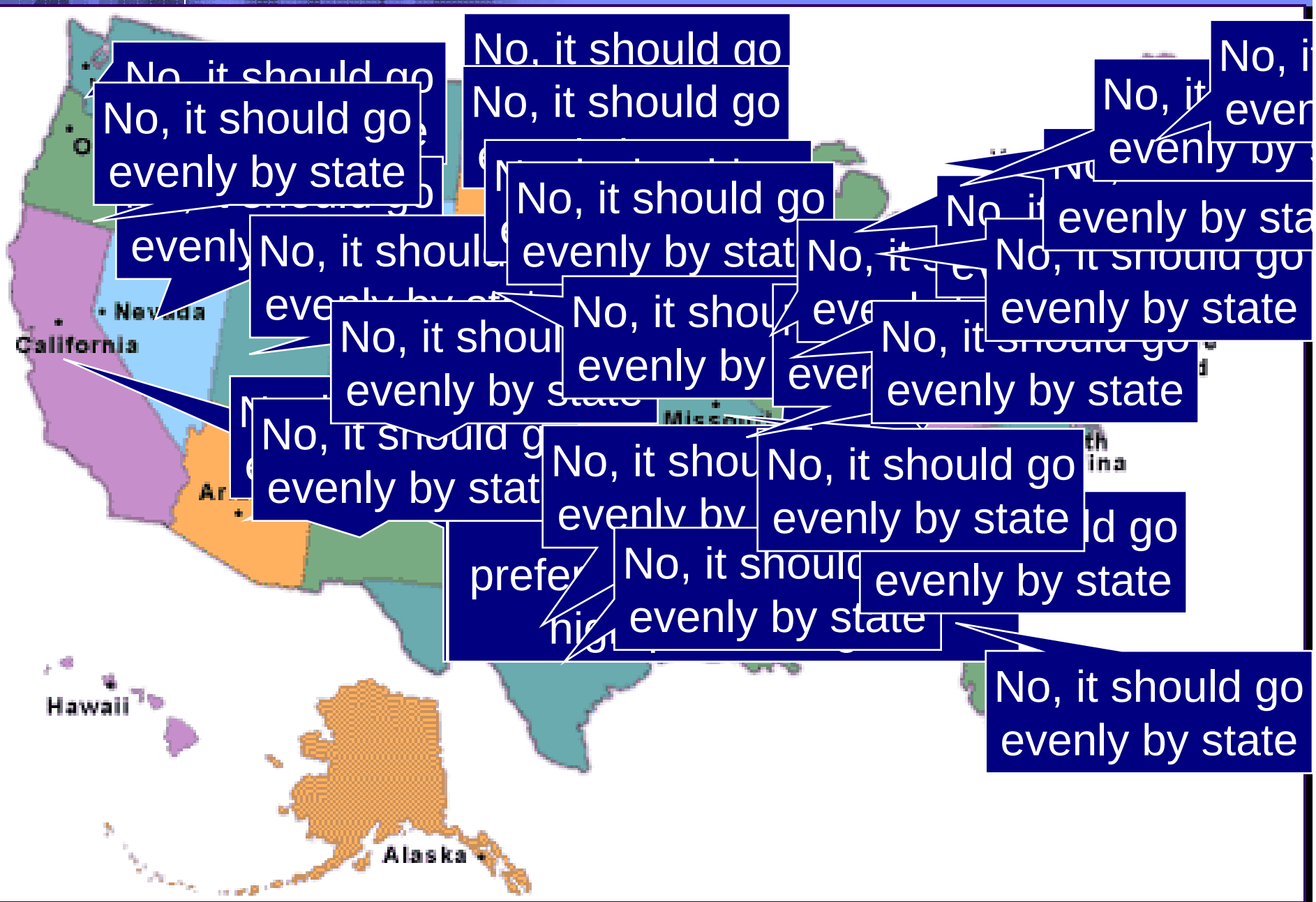
But!!

There is a better way

How?

Emulate the model we use for funding of

Homeland Security



No, it should go

No, it should go

No, it should go

No, it should go evenly by state

evenly

No, it shoul evenly by st

No, it should go evenly by stat

No, it shou evenly by

No, it shoul evenly by state

No, it should g evenly by stat

No, it shou evenly by

prefer hig

No, it should evenly by state

No, it should go evenly by state

ld go evenly by state

No, it

ever

No, it evenly by state

No, it should go evenly by state

No, it evenly by sta

No, it should go evenly by state

No, it should go evenly by state

No, it should go evenly by state

Which means…

We should split the $6B system across

every congressional district in the country

How do we do that?

Congressional districts are sized by population
(1 Representative per N people).

What else scales with the number of people?

Schools.

Of every N people a certain percentage are children, & kids need schools – so every congressional district has schools

So what do we do?

- Put part of the $6B machine in every high school and every post-secondary institution (colleges and universities) in the country

So how does this work?

- There are about 35,000 high schools, and 12,000 colleges & universities in the country - ~47,000 in all
- According to the 2001-2002 data (slightly old): http://nces.ed.gov/programs/digest/d03/tables/dt005.asp
  - 22,180 public high schools
  - 2,538 private high schools,
  - 8,155 combined elementary/secondary private schools
  - 2,245 public post-secondary institutions
  - 2,777 private not-for-profit post-secondary institutions
  - 4,463 private-for-profit post-secondary institutions
- 6M/47,000 = 127.6 -- so this allows for a 128-processor cluster, fully loaded, for every school in the country

  – Again – 128 CPUs/school * ~47,000 schools = ~6M CPUs

So how does this work, again?

- You'd give a 128-way cluster, fully loaded to every high school and post-secondary school in the country.

- Each cluster would *also* have a big display system – wall-sized, tiled or HDTV – so that it would go in the school's auditorium or conference room, as well as webcams, for distributed group-to-group collaboration (Access Grid model)

So how does this work, again? – The networks

- School Clusters would be very tightly linked to neighboring school clusters, across direct 1500 nm links
  - Each school cluster would have 16 (10+10) Gbps links, to connect to other schools in the area.
  - All <100km technology, no WAN/SONET needed, since every high school is within 100km of another high school

So how does this work, again?

- Time would be split among various uses:
  - At night: TeraGrid model
    - Country-wide calculations allocated by some central authority – DOE or NSF
  - During day: Education
    - Distributed learning in classes (not just individuals), team-teaching across different states, …
  - In the evening: Entertainment
    - multi-player inter-school video game contests
      - our Madden football team against the Madden footbal team across town
    - Movies, TV, Class projects (1-person movies)

How to build it, Take 2: Re-allocate the money

- ~~$2B~~ $1B worth of DRAM
  - In 2010-2011, this will be ~24 PB, or 4GB/processor
  - Equivalent to 24 million DIMMs
- ~~$1.5B~~ $3.0B worth of optics for 100 kilometer links
  - e.g., (1+1)-lambda XCVRs, 1500 nm SMF, 10Gbps / lambda
  - At ~$4600 / lambda, this allows 750K lambdas: 16 per school cluster
- ~~$1.5B~~ $1.0B of storage (hard disks & RAM disk cache)
  - Equiv to ~2 disks, $125/disk, per processor core: ~4 PB total
- $1B for Switches, Boards, Power supplies, DC/DC Converters, Racks, Cooling, ~~Data center, Building for the data center~~, ~~power plant~~ to ~~power the system,…~~, and 6M processor cores, …., and 47,000 big display/monitor systems, plus ~200,000 web cameras

There are a few advantages to this picture

- Solves the power & cooling problem
  - Evenly distributed across the whole country, negligible extra load to any particular part of the power grid
- Solve management problem - schools *will* manage the systems when used for education &entertainment
  - The system administrators will hear from the other students: “We have a game of video football against the high school across town on Friday – you *must* have the system up and running.”
- Huge benefit to education in HPC/HEC & Networking
  - Ubiquitous parallel processing - every kid will have a 128-way cluster in his school to work and play with, that will be directly tied to neighboring schools and across the country.

The Six Million Processor System...
...Through Volunteer Computing

Carl Christensen

University of Oxford

Department of
Atmospheric Physics

carlc@atm.ox.ac.uk

Volunteer Computing

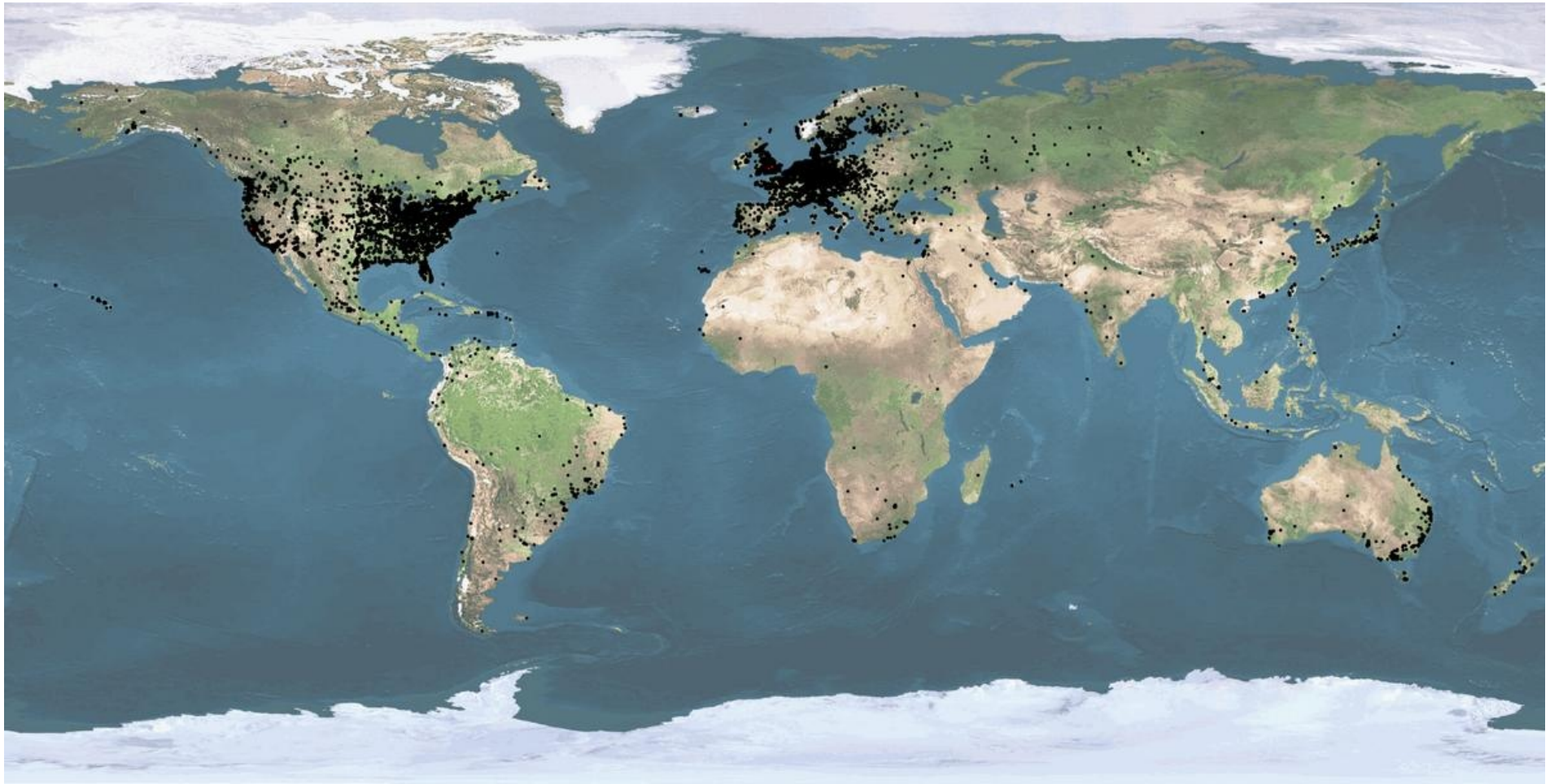
- Previously called “public resource distributed computing”, often confused with “plain” grid or distributed computing
- A specialized form of “distributed computing”
- Was around before '99 but really took off with SETI@home project
- S@H with 500K users ~1PF = 1000TF
- Earth Simulator in Yokohama = 40TF peak
- CPDN running at about 60-70 TF (30K users each 2GF machine average, i.e. PIV 2GHz)
- Best benefit – performance & price

SC|05

Volunteer Computing Potential

- SETI@home - Has had 5.5 million users (total), 500K concurrently (typically)
- On order of a 6 million processor system already!
- AOL – 150 million users (*Newsweek*, 09/30/02)
- 75% of Americans have Internet access (*NW*, 10/11/04)
- 934 million Internet users worldwide (2004, *Computer Industry Almanac*)
- Estimated 1.21 billion PCs worldwide (2006 projection, *Computer Industry Almanac*)
- 0.5% of worldwide PCs in 2006 = 6 million processor system!

climate*prediction*.net BOINC Users Worldwide
>100,000 users total: ~30,000 at any one time (trickling)



As CPDN Principal Investigator Myles Allen likes to say...

"this is the world's largest climate modelling supercomputer"

Berkeley Open Infrastructure for Network Computing (BOINC)

- http://boinc.berkeley.edu
- An open-source (GNU LGPL), multi-platform capable vertical application for volunteer computing
- Used by SETI@home, climate*prediction*.net, Predictor@home,Einstein@home, Rosetta@home, LHC@home and others, with more coming!
- Offers a complete client and server-side API to get an application developed cross-platform, as well as an OpenGL graphics engine for screensaver etc
- ports the “tried, true, and tested” SETI@home infrastructure for use by anyone
- Funded by the US National Science Foundation

BOINC Benefits

- Open-source “free” software
- Multi-platform support (Windows, Mac OS X, Linux, Sun Solaris, pretty much anything supporting GNU tools e.g. autoconf/make, g++, etc)
- A complete “vertical application” for volunteer computing – client software API, server, website etc
- Small staff (typically 2-6) required for development and support of what is basically a large supercomputer
- Distributes the power consumption, hardware & software maintenance (“distributed sysadmin”)
- Not just for volunteer computing – also useful (and used) on corporate Intranets, research labs, etc.

Volunteer Computing Caveats

- Greater security concerns
  - Result falsification => redundant computing by users, result validation by the project
  - Website/server hacking => email validation, server patches, upload certificates
  - Malicious executable distribution => code-signing key pair
  - Ref: http://boinc.berkeley.edu/security.php
- User attraction and retention
  - With BOINC doing most of the work, VC becomes more of a marketing issue than a technical issue

CPDN / BOINC Server Setup

- Our “supercomputer” hardware is basically a few “off-the-shelf” servers, about £10K ($17K) total.
  - Database Server – Dell PowerEdge 6850, two Xeon 2.4GHz CPUs, 3GB RAM, 70GB SCSI RAID10 array
  - Scheduler/Web Server – Dell PowerEdge 6850, two Xeon 2.4GHz CPU, 1GB RAM, also usually <<1%
  - Upload Servers – federated worldwide, donated, so vary from “off the shelf” PCs to shared space on a large Linux cluster.

BOINC!

SC|05

Public Education via Volunteer Computing

- CPDN has public education via the website, media, and schools as an important facet of the project
- Website has much information on climate change and related topics to the CPDN program.
- Schools are running CPDN and comparing results, especially during National Science Week (starts 12/3/04) with special events at U Reading
- Students will host a debate on climate change issues, compare and contrast their results etc.

Students at Gosford Hill School, Oxon viewing their CPDN model

- Currently focused on UK schools, but as projects added and staff resources are gained plan to expand to other European schools and US schools

Does SIZE Matter?

Satoshi Matsuoka
Professor
GSIC Center, Tokyo Institute of Technology
and
Co-Leader, NAREGI Project, National Institute of Informatics

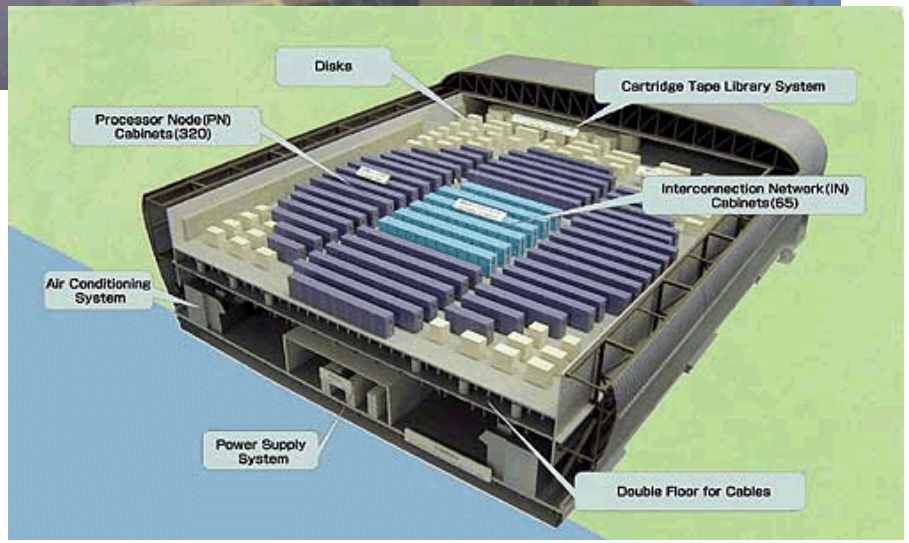
What SIZE matters for a 6 million PE machine (i.e., impediments)

- Physical Machine SIZE
- Network SIZE (Dimension, Arity, etc.)
- PE SIZE (#CPUs)
- Memory SIZE
- Node SIZE (#PE, Mem, BW, Physical, Watts)
- Chip SIZE (#PE, Die Area)
- Linpack SIZE (Rmax)

- What we brag about being BIGGER: (typically) Physical Machine SIZE, PE SIZE, Linpack SIZE, …
  - Human Nature?
  - What becomes an impediment?
  - What does really matter?

Physical Machine SIZE & Physical Node SIZE

- Japanese Earth Simulator (2002)
- 640 NEC SX-6 modified
  - 5120 Vector CPUs
  - 8 CPUs/node, 2 nodes/rack
  - 320 CPU + 65 network racks
- 40TeraFlops (peak), 36TeraFlops (Linpack)
- 7-8 MegaWatts
- $400-$600 million
- Size of a large concert hall (3000 sq. meters)
- Can we build anything substantially bigger? (Like Stadiums and Beyond)
  - SX-8 has same flops density, x2 power density



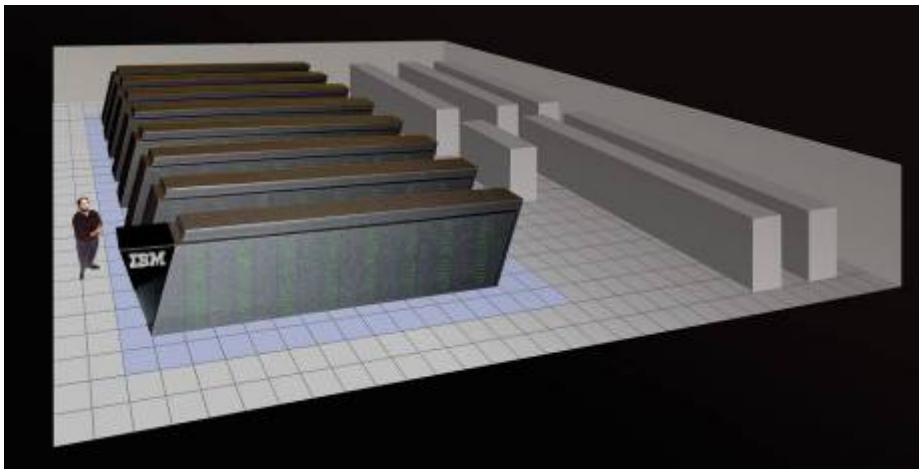
(Earth Simulator Picture from JAERI web

The Pentagon --- the SIZE Limit

- Still largest building in the world w.r.t. floorspace, with floor area of approx. 600,000 sq. meters
- x200 ES floor space => could fit a million CPUs only
- So SIZE matters

- 8 Petaflops, ~1000MW
  - 1/2 Hoover Dam
- This is foolish, as we know
  - But just to be sure we recognize it is foolish

PE and Node SIZEs

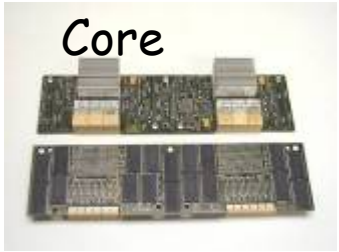


360TeraFlop, 220 sq. meters, 1.5MW

System
(64 cabinets, 64x32x32)

Cabinet
(32 Node boards, 8x8x16)

Custom PowerPC440 Core



Node Board
(32 chips, 4x4x2)
16 Compute Cards

Compute Card
(2 chips, 2x1x1)

Chip
(2 processors)

2.8/5.6 GF/s
4 MB

5.6/11.2 GF/s
0.5 GB DDR

90/180 GF/s
8 GB DDR

2.9/5.7 TF/s
256 GB DDR

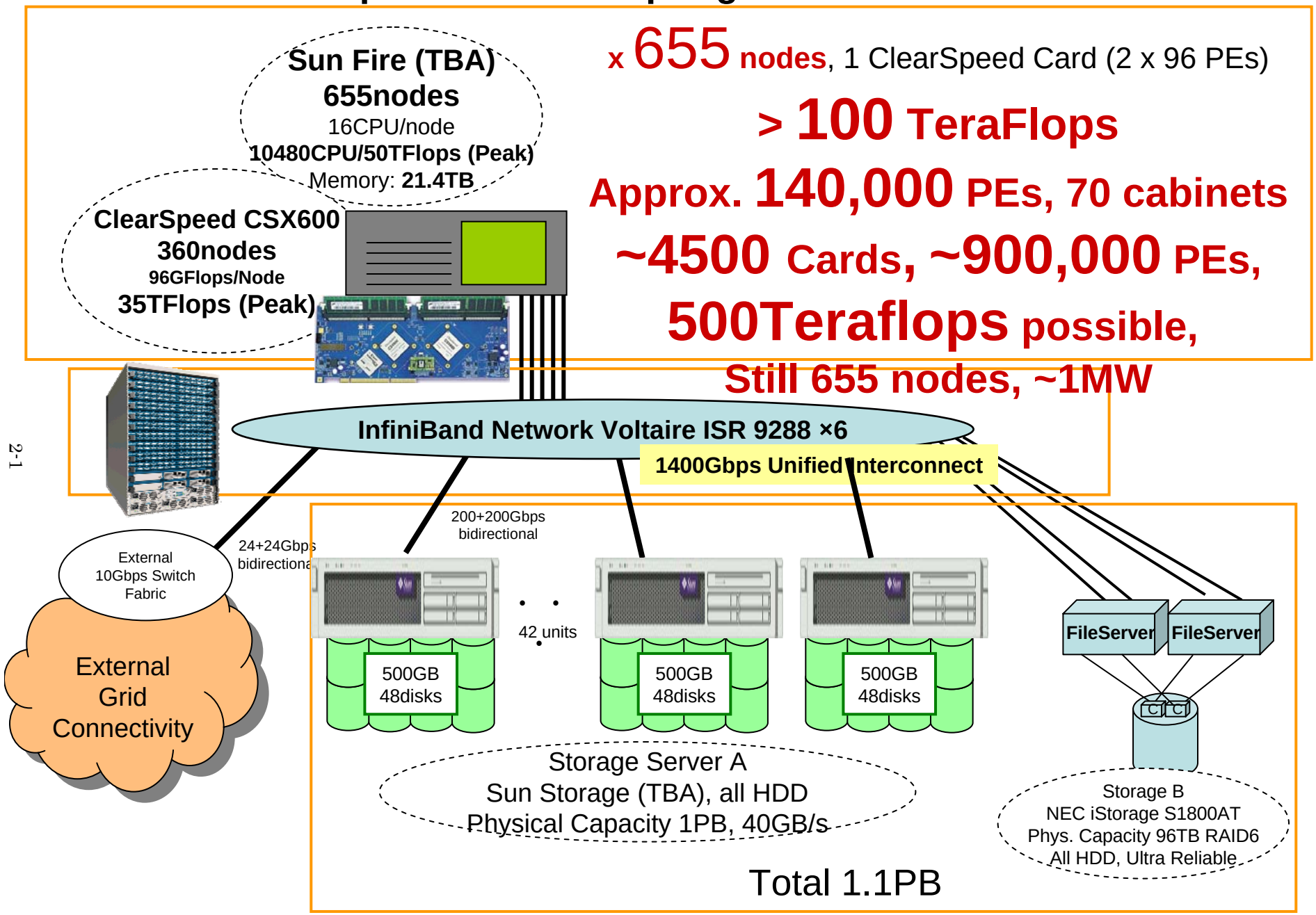
~260,000 PEs
180/360 TF/s
16 TB DDR

2048CPUs/rack
25KW, 5.7TF

No-Brainer SIZE Scaling

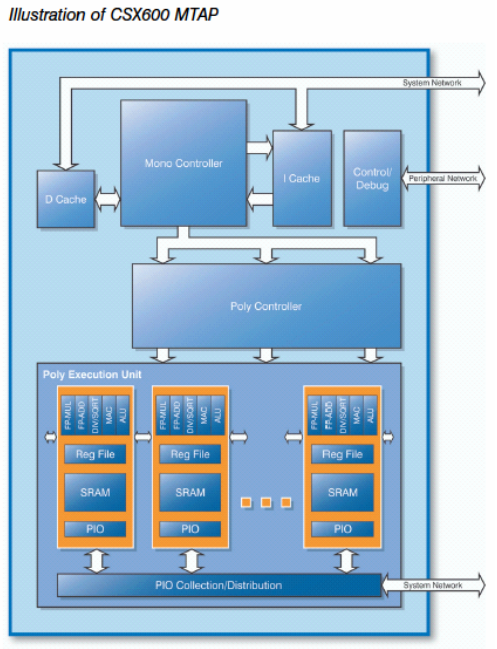
- 6 million processors = 3 million nodes
  = 3000 BG/L racks
  = x 46 current size = ~10,000 sq. meters
  = 17 Petaflops
  Doubling of cores => 1500 racks, ~5,000 sq. meters
  => just x2~3 Earth Simulator
  - Physical Machine SIZE somewhat matters
  - PE and Node SIZE may matter a lot
    - Flat machine abstraction no longer may work
- 30KW/rack => ~100MWatts
  - A small nuclear power plant
  - Power SIZE somewhat matters

NEC/Sun Campus Supercomputing Grid Core Infrastructure @ Titech GSIC - to become operational late Spring 2006 -



Does any SIZE Matter?

- x7 scaling achieves 6 mil PEs
  = ~2000 sq. meters
  = 7 MWs
  = 3.5 PetaFlops
  Only 4500 Opteron nodes!
  ~30,000 Cards, 60,000 Chips
- x4 PE increase with .65 nm
  - Just reduce the SIZE (# cards)
  - Only 1000 nodes, 7000 cards
  - 500 sq. meters, ~2MWs
- 6 million PEs achieved!
  - Overall system much tractable
- **SIZE** does not matter



96 SIMD PEs/Chip

So What Matters?

- It is not the SIZE, but the HIERACHY of SIZEs and their ABSTRACTIONS
  - (w/appropriate stuff in the right place in the hierarchy)
- Fundamental CS (in fact Engineering) discipline
  - O(log n) instead of O(n)

- In building large systems, how often we forget this…
  - We get macho with full crossbars, +100W single threaded processors, full BW vector loads, flat address space…
    - As a result, SIZE will matter
  - NASA Columbia and Riken Combined Cluster being the only "hierachical" systems in the Top100/500
    - Should always be thinking of ways where SIZE does not matter (O(log n), O(sqrt n), etc.)

HPC - 6M Systems - Is there Any Hope?

Applying New Parallel HW/Programming Paradigms to Mission Critical HPC Applications

SC'05
Nov 18, 2005

Jim Taft
jtaft@nas.nasa.gov

NASA

NAS
NASA Advanced Supercomputing

First - There are no more Seymour Crays

Not sure why, but no one comes close to his long term architectural brilliance

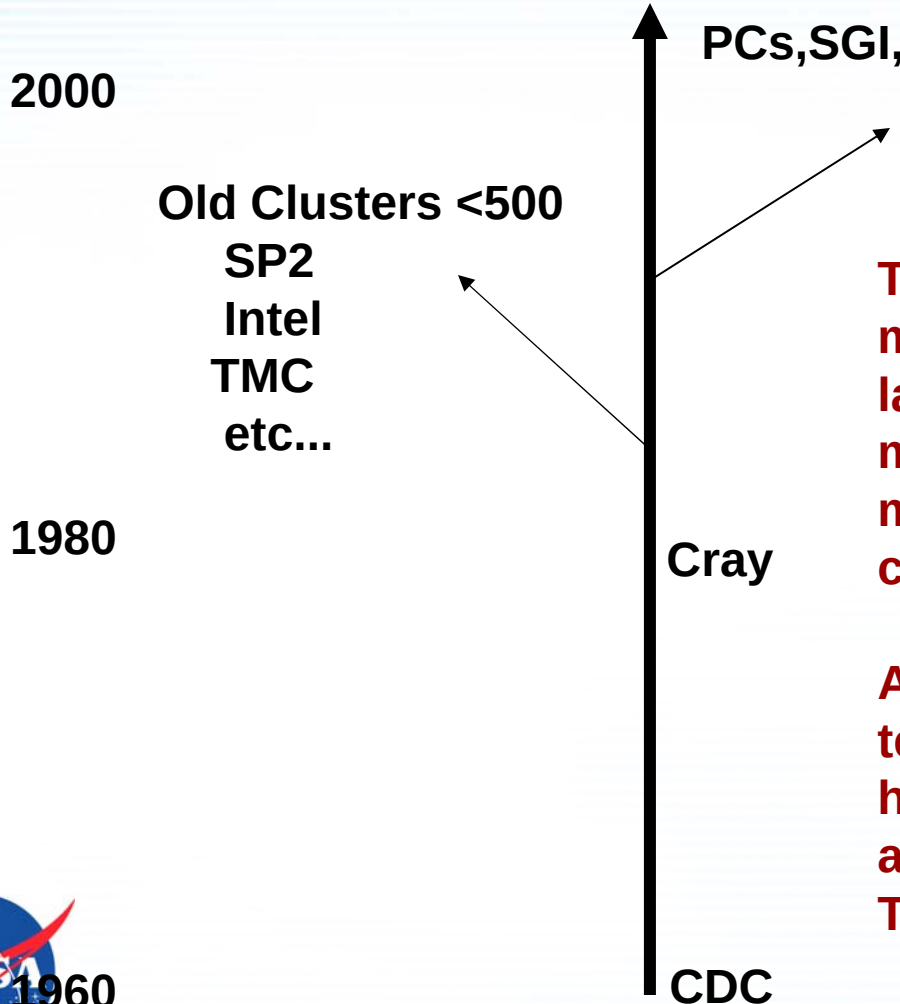
Steve Wallach is perhaps a number two that gets it - but he is out of the biz

Not much else to choose from

Why the deselect in this business:

- *Wall street?*
- *Loss of Venture Capital?*
- *Dumbing down of the community - Do we really mentor/teach HPC now?*

Family Tree of Supercomputing (2001)



There are > 100,000,000 shared memory systems. There are <100 large HPC clusters. Why should we massively change our programming model for this tiny pittance of clustered systems?

Answer: We shouldn't. There is no technical reason whatsoever. We have simply let the vendors get away with it. It's time to stop. 100 TFLOP/s SSI is around the corner

NAS
NASA Advanced Supercomputing

Is There any Hope - The Pessimist's View

- **Assertion: Real Science + Computer Science Often Equals No Science**
  - In the not so distant past many orgs committed major funding in exotics
  - Dozens of expensive, time consuming, non-performing, dead ends.
  - Complex, expensive, poor system software support, admin nightmare

- Current glitzy CS projects are diverting meager HPC HW/SW eng resources
  - Infatuation with Add in FPGAs, acccelerators, etc - good for very few
  - Virtually useless for general population

- ASCI did nothing but buy COTS - Rate of Acceleration = 0
  - Margins were so meager no R&D benefits ensued to vendors
  - No serious architectural wins were realized in either SW or HW
  - Results - serious HPC providers are on the edge

6M System - Some Observations

- ***Facts:***
  - ***It will be exotic - a bastard child no matter who the vendor***
  - ***It will suffer from massive system reliability and system SW issues***
  - ***It will likely come on line years after original target date***
  - ***It will have extremely limited utility and applicability***

- ***Some Ancient History:***
  - ***PIM, Systolic arrays, custom ASICS, accelerator boards are decades old***
  - ***If they had any reasonable utility - they would be in abundance***
  - ***Are we doomed to repeat history with just a new layer of glitz***

6M - Should We Do it? - No

Fact: A very few years ago - 3 TFLOP/s was "it" at the labs for >100M

This system was a shared resource for an entire National Lab.

This was followed by 10 and 30 TFLOP systems for 100Ms more.

Today: You can buy 3 TFLOP/s SSI systems for about 4M

Supports higher percentage of peak, vastly easier to use

Tiny footprint and trivial to maintain, highly reliable

BTW - Received no major government funding awards

12 TFLOP/s SSI is available next year - 12M?

Observation: Let's give everyone an ASCI Blue/White for their birthday.

Rate of scientific discovery would be explosive - envigorating

Could put focus back on the science - not computer science

Ok, I give up - Let's build it - Optimist's View

- Where are you Gene Amdahl? Yikes
  - Well I guess it scales for some problem(s)
  - Is it real science or are we just kidding ourselves?
  - Quantum mechanics at work - ask for an answer - get no performance
  - The Star 100 problem to the N th power?

- Let's see: Hmmmm we need:
  - A new programming language?
  - New debuggers, analyzers, etc
  - New OS kernels?
  - Vastly improved I/O sub-system?

- It Comes online when?
  - Ok it's just a few years out
  - A few small systems are distributed
  - Ooops - we left out some essential communications - we fix - retry
  - IBM's Power X has caught up?

How about learning from the Past?

- *We constantly ignore lessons learned*
  - *SSI is simpler in all regards*
  - *SSI is scalable to relatively large size*

- *The programming models of the classic CDC7600/Cray have been dropped*

- *A "sea" of CPUs SSI system is classic and high performance*
  - *Large CPU count (>10K) with modest local memory*
  - *Large global shared bulk memory*
  - *User control of block transfers to local memory*

- *Just happens to look much like CDC7600*
  - *Achieves high percentage of peak on traditionally tough CFD problems*

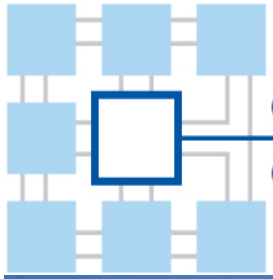
Summary/Observations

- *Money spent on such architectures prematurely inhibits scientific progress*
  - *Buy a series of useful modestly powerful systems - you'll be way ahead*
  - *Wait a couple of years the industry will catch up - Ala "blue" systems now*

- *It has virtually always been the case that such exotics don't work*
  - *They are also eclipsed a few years later by mainstream offerings*
  - *Is it that important to divert precious people/budgets with marginal return*

- *100 TFLOP/s and 100 Tbyte SSI is around the corner - 3-4 years.*
  - *Other 100 TFLOP exotics will still be struggling to meet their ambitions*

- *Funding orgs should back off and concentrate on usefulness and productivity*
  - *Spread the wealth philosophy has done NASA well - Columbia*



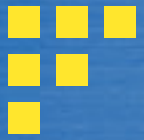
CENTER FOR HIGH-END COMPUTING SYSTEMS

Virginia Tech

The 6 million processor system

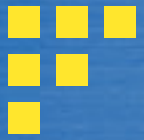
Dr. Srinidhi Varadarajan
Director
Center for High-End Computing Systems
Virginia Tech

SC 2005



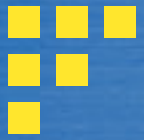
Challenges

- Power
- Cooling
- Computing Model
- Communication
- Extreme scalability
- Usability
- Reliability
- Floor space



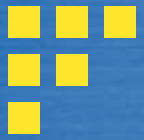
Architectures

- Extension of BlueGene style low power architectures
- Constellations of multi-core systems
- Traditional clusters
- Everything else including clustered toasters



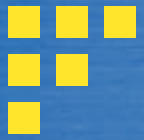
Perspectives

- Look back at processors in the pre-VLSI era
- All internal functional units were independent components
- Programming was structure superimposed over a resource allocation problem



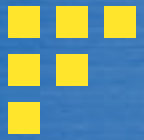
Thoughts over a beer

- Consider *a highly* superscalar processor
  - 6 million functional units
- More complex functional units
  - Linear algebra units
  - FFT, convolutions
  - Image processing
  - …
- Redundancy through large numbers of identical functional units.



Thoughts over the second beer

- Programming Model: Combination of Von Neumann and dataflow.
- High levels of integration may yield a system that fits within the power budget.



Issues

- Memory model: Globally addressed memory would place phenomenal bandwidth/latency requirements.
  - Computation in memory?
  - Local memory with message passing?
- Usability: Can such a system be made usable?