Green HPC: A System Design Approach to Energy-Efficient Datacenters

By

Kurt Keville

B.S. General Engineering, 1983

United States Military Academy

Submitted to the System Design and Management Program
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering and Management
at the
Massachusetts Institute of Technology
May 2011
[June 2011]

© Kurt Keville. All Rights Reserved.

ARCHIVES

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

JUL 20 2011

LIBRARIES

The author hereby grants to MIT to permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author: ______________________________
Kurt L. Keville
System Design and Management Program
May 2011

Certified By: ______________________________
Stephen R. Connors
Director, Analysis Group for Regional Energy Alternatives
Thesis Supervisor

Accepted By: ______________________________
Pat Hale
Director, System Design and Management Program

Green HPC: A System Design Approach to Energy-Efficient Datacenters

Submitted to the System Design and Management Program on May 6, 2011 in partial fulfillment of the requirements for the degree of Master of Science in Engineering and Management.

Abstract

Green HPC is the new standard for High Performance Computing (HPC). This has now become the primary interest among HPC researchers because of a renewed emphasis on Total Cost of Ownership (TCO) and the pursuit of higher performance. Quite simply, the cost of operating modern HPC equipment can rapidly outstrip the cost of acquisition. This phenomenon is recent and can be traced to the inadequacies in modern CPU and Datacenter systems design. This thesis analyzes the problem in its entirety and describe best practice fixes to solve the problems of energy-inefficient HPC.

Thesis Supervisor: Stephen R. Connors

Title: Director, Analysis Group for Regional Energy Alternatives

Acknowledgements

I would like to thank the many classmates, faculty, staff, and others that have contributed to my work on this thesis and to my experience in the MIT System Design and Management Program, especially all of my colleagues in the SDM cohorts of 2009, 2010, and 2011, who shared their experience and knowledge.

To the students and faculty I have met through my assorted Energy and HPC related projects whose dedication to the search for alternative approaches to building the processing power necessary for University, DOE National Lab, and DOD HPC facility level research and development helped motivate me in my own work. In particular, the ad hoc MIT HPC Task Force, headed by Chris Hill of MIT EAPS, has been a rich resource from which I could cull and contribute considerable quality material.

To Pat Hale and the rest of the SDM staff for continually improving the SDM program, making it an excellent example of how Engineering and Management education should be taught.

To Steve Connors, who provided encouragement, direction, feedback, and support through the process of this work.

Finally, to my wife Lauren, for her considerable help in motivating me through the difficult periods in the SDM program.

Table of Contents

Conventions

For shorthand purposes DC represents Direct Current while the acronym DC, used often in the literature for Datacenter, is not used here. Likewise, AC will stand for Alternating Current while Air Conditioning will be the abbreviation HVAC or spelled out to avoid confusion.

List of Figures

Chap. 1: The Power Problem in Modern Datacenters

Impetus and Motivation

DARPA's 2008 exascale report forecasts a greater than 100 megawatt (MW) power budget for a 1 exaflops machine if current trends continue (Hemmert, 2010). This is not a viable strategy and therefore calls for a thorough investigation into a redesign with energy-efficiency and Green HPC as its foundation. As we will see, HPC Datacenters are a superset of the broader Datacenter design and we can learn a considerable amount from the way the data provisioning business community has improved on the archetypal design. Additionally, the DOE, who maintains a plurality of the Federally-funded HPC Datacenters in the US, is mandated by Executive Order 13423 to reduce facility energy intensity 30% by 2015 (Energetics Incorporated, 2010).

> *"The push for exascale computing will provide a driver for an unprecedented level of energy efficiency."*
> *—Scott Hemmert (Hemmert, 2010)*

In this study, we will break up an analysis of the HPC Datacenter by looking at its constituent parts. These parts will be further subdivided so that we can look at the components that feed into these subsystems. We will do this with an eye towards calculating and analyzing the energy usage of these subsystems as well as determining various energy efficiency improvement approaches. We will pay proportionately appropriate attention to the "power hogs" to reduce their relative energy use.

We are defining a Datacenter as a conditioned room within a building that houses all of the computational machinery used in either a cluster or an MPP, as defined and exemplified by the Top500 list (and derivatively, the Green500 list). All but 3 of the Top500 list supercomputers are considered Scalar Processor Architectures, and all but 2 are either traditional computer clusters or MPPs. We will therefore consider HPC (or a broad definition of a supercomputer) to be one of these architectures. We will only just briefly address the external building concerns that impact energy efficiency since the bulk of our discussion will concern the computer room in the building that houses this Datacenter. This is where most of the power is being used and therefore needs to be the focus of our analysis. Datacenters are nominally built out of a large number of 42U racks (where a U is a standard Datacomm industry measurement of 1.75 inches, usually the height of a server) populated with server and storage equipment, often in 1U, 2U, or 4U cases, and are cooled with a large HVAC subsystem.

What this covers

This is both a descriptive and prescriptive assessment of what will happen in the HPC business this decade with particular attention to the "Path to Exascale" as defined in the DARPA report. Following the guidelines specified herein will pay dividends in contemporary design and will lay out groundwork to define an approach to future exascale systems. Datacenter IT equipment usually has a 3-year use and then needs to be retooled. This may come in the form of component swaps, such as upgrading to the next generation CPU by pulling the older processor from its socket and replacing it with the latest model, or it could come in the form of a replacement of the entire rack. The infrastructure is cost-effective for 10 years and then it, too, is reassessed and retooled or rebuilt, historically. It won't be scrapped, but it will be cheaper to build another center than to continue using it relative to starting fresh at a new site to run better and faster HPC applications with the then predictably improved hardware.

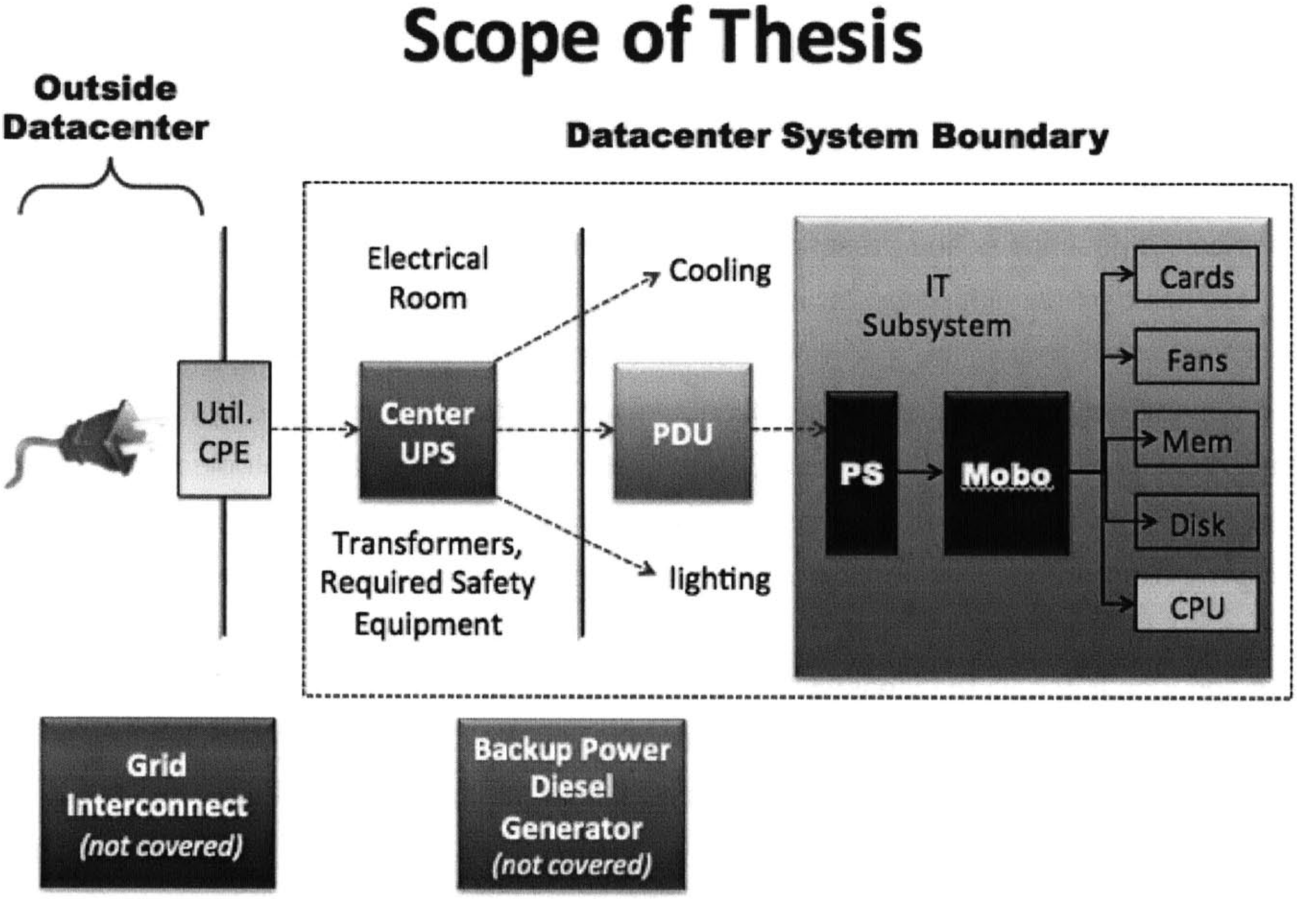

Scope of Thesis
Outside Datacenter
Datacenter System Boundary
Util. CPE
Electrical Room
Center UPS
Transformers, Required Safety Equipment
Cooling
PDU
lighting
IT Subsystem
PS
Mobo
Cards
Fans
Mem
Disk
CPU
Grid Interconnect (not covered)
Backup Power Diesel Generator (not covered)

Figure 1: The Datacenter is a System of Systems

What this doesn't cover.

We will not consider reviving old technologies or reuse of old equipment in an attempt to recapture some of the embodied energy represented in the construction of that equipment. While we are able to draw some of the "Reduce, Reuse, Recycle" methods from the sustainable construction industry, this is impractical to implement in the modern Datacenter design models. We will also not discuss improvements you can make outside of your Datacenter including geographical load-balancing, "chasing the sun," or "chasing the moon" strategies. It is safe to say that these strategies are available outside of a discussion of a new construction or operational recommendations. To that end, this paper does not address methods of retrofitting existing centers. Additionally, this doesn't cover topics subsumed under the larger title of "Grid Computing" since that captures improvements at individual Datacenters.

Target Audience

While accessible to all audiences, this paper is specifically targeted at the Datacenter designers and stakeholders that need to come together to create a comprehensive construction and operations plan for a new HPC facility. This includes the facilities engineers that are responsible for the detailed schematics and blueprints of the building as well as members of the community who have an interest in the reduced power use and the resultant carbon dioxide emissions associated with the center. In addition to those on the planning side, there are guidelines specified here that are applicable to operations personnel, in particular systems administrators and application coders (the HPC users).

The Problem

The world and the US, in particular, is very inefficient in transporting power from where it is generated to where it gets consumed. There are inefficiencies at every step of the electrical transport and computational chain.

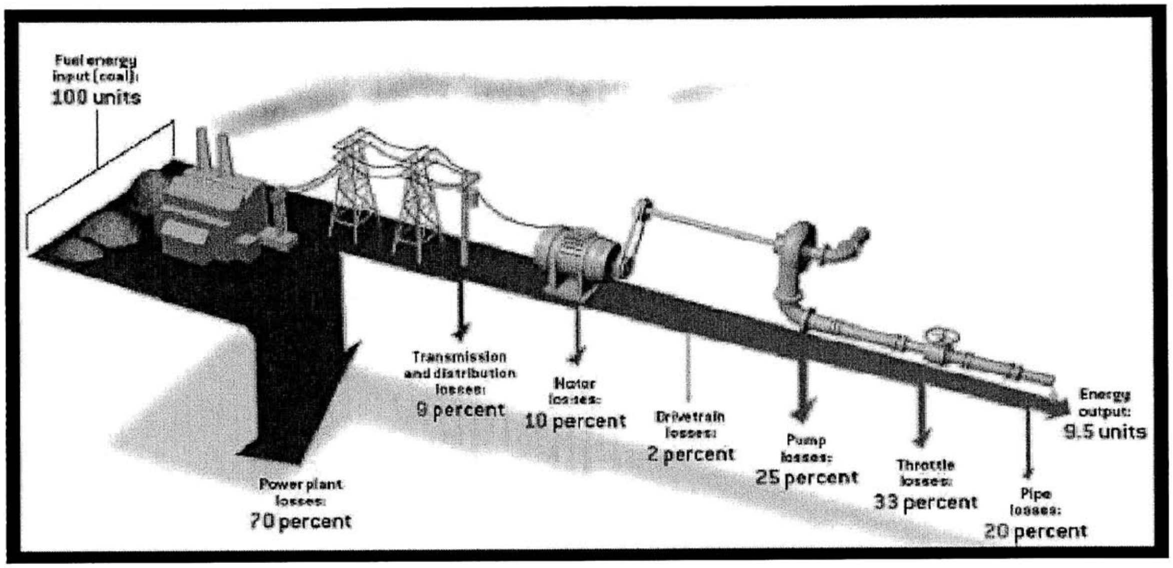

Fuel energy input (coal): 100 units
Power plant losses: 70 percent
Transmission and distribution losses: 9 percent
Motor losses: 10 percent
Drivetrain losses: 2 percent
Pump losses: 25 percent
Throttle losses: 33 percent
Pipe losses: 20 percent
Energy output: 9.5 units

Figure 2: Inefficiencies in power delivery systems (Lovins, 2008)

Once the power gets there we use it very inefficiently within the building.

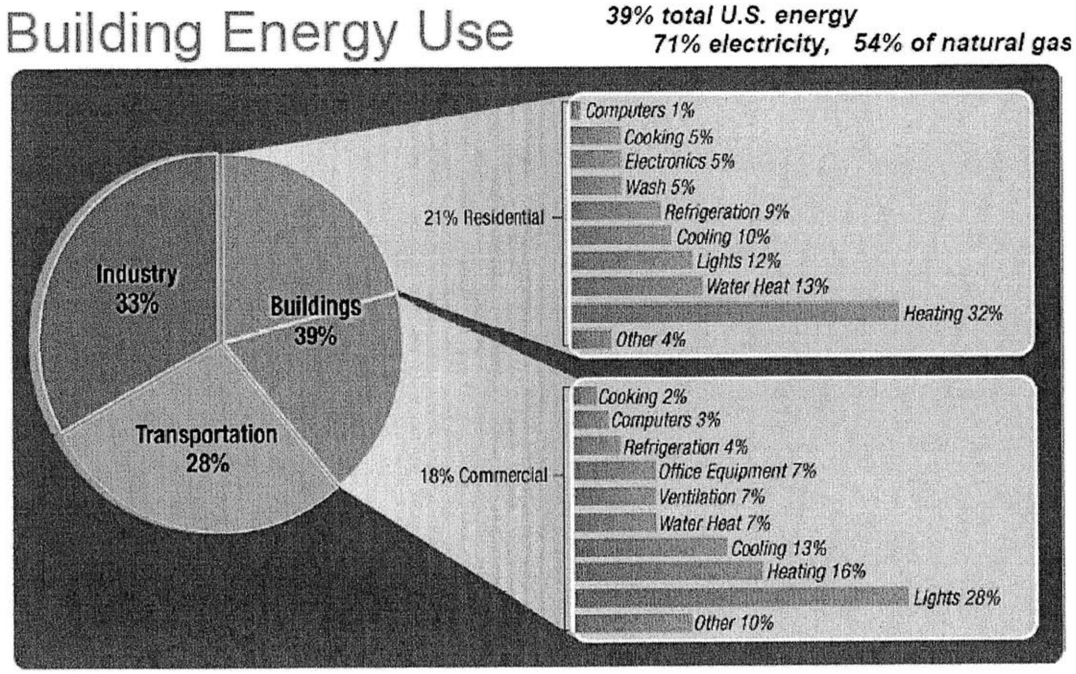


Figure 3: Allocation of Energy Use in the US (Michaels, 2009)

And lastly, we use power very inefficiently on the subsystems components within the building.

Typical power use in a Datacenter

Cooling and IT have the majority of power "ownership" in the Datacenter.

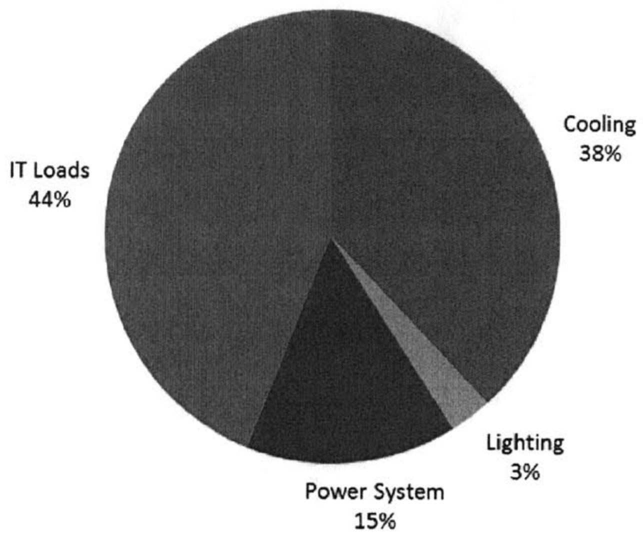
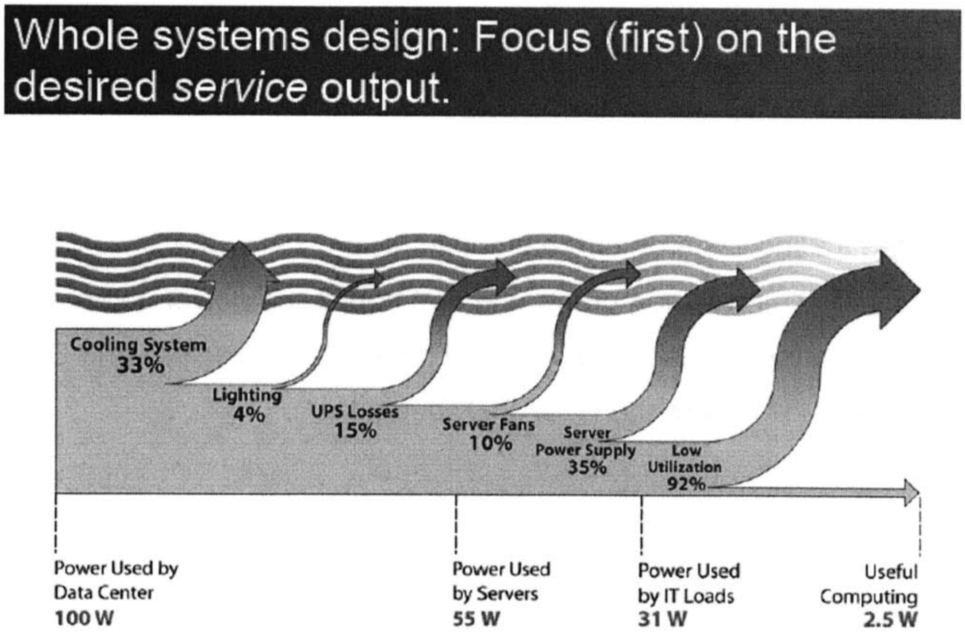


Figure 4 (Rasmussen, 2003)

But losses are not distributed evenly. Cooling is around 67% efficient and most servers have very poor aggregate efficiencies as demonstrated below.

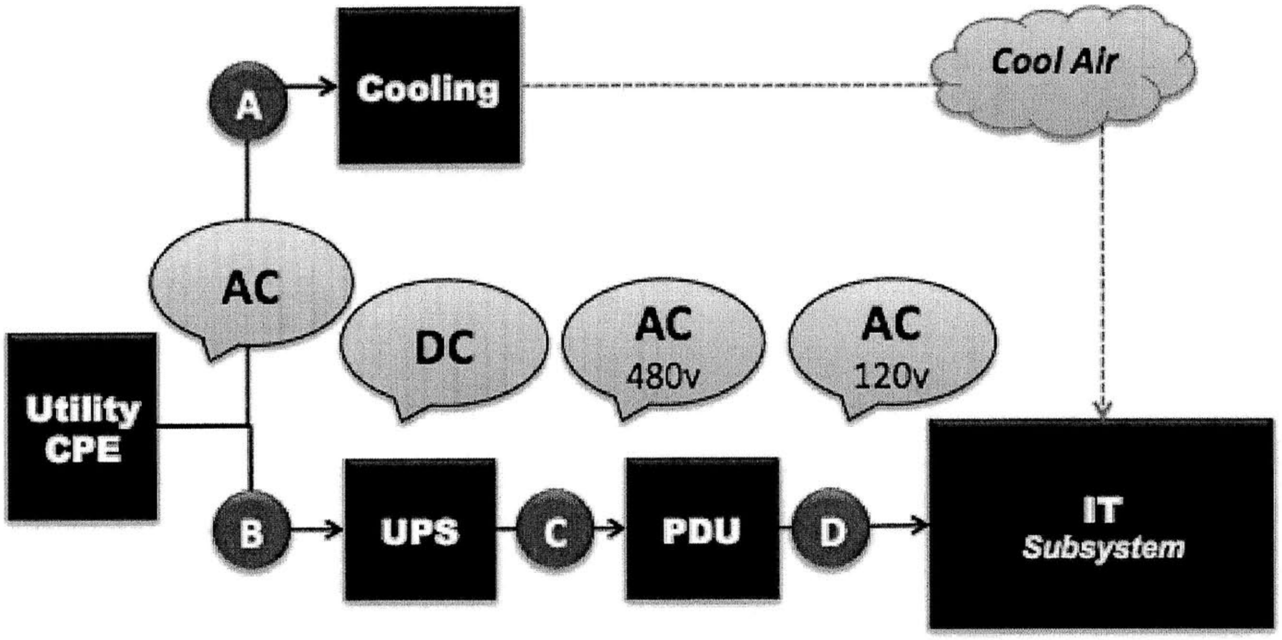


Whole systems design: Focus (first) on the desired *service* output.

Figure 4: Power usage diagram (Newman & Palmintier, 2008)

Calculating Efficiency Today (PUE)

Traditional Datacenter Layout



$$PUE = \frac{A + B}{D}$$

Figure 5 A typical “chain of inefficiency” in a Datacenter

The problems at the trailing edge of our chain of inefficiency are where we need to make appropriate fixes since they will then propagate up through the power chain and yield us significant savings. The task ahead is to drive the cost of Infrastructure and Energy VAREX down to equivalency or lower with the IT CAPEX budget.

Annual Amortized Costs in the Datacenter for a 1U server.

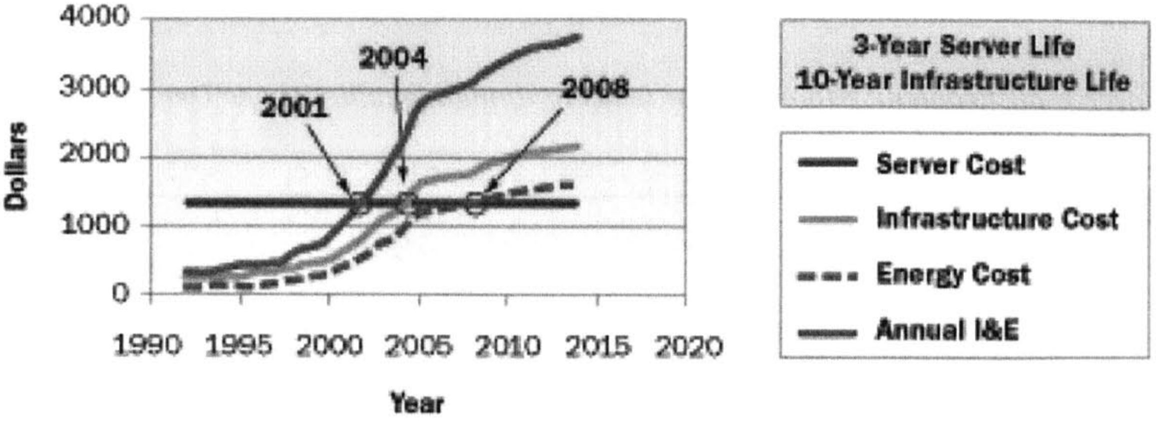


Figure 6: “Infrastructure and Energy Cost (I&E) will be 75% of the cost in 2014 and IT will be only 25%.” (Belady, 2007)

Harvey Michaels, Research Scientist at the MIT Energy Initiative, says the way to get energy savings is through energy efficiency, equipment improvements, and automated operational changes such as those espoused in the discussion of our programming methods. You can't count on consumers to substantially change their behaviors in the pursuit of energy conservation so you have to build in energy savings into the equipment. The behavior modifications we espouse will primarily be targeted at the HPC Center employees, whose job it is to liaise with the customer and adapt their application code to the Green HPC model.

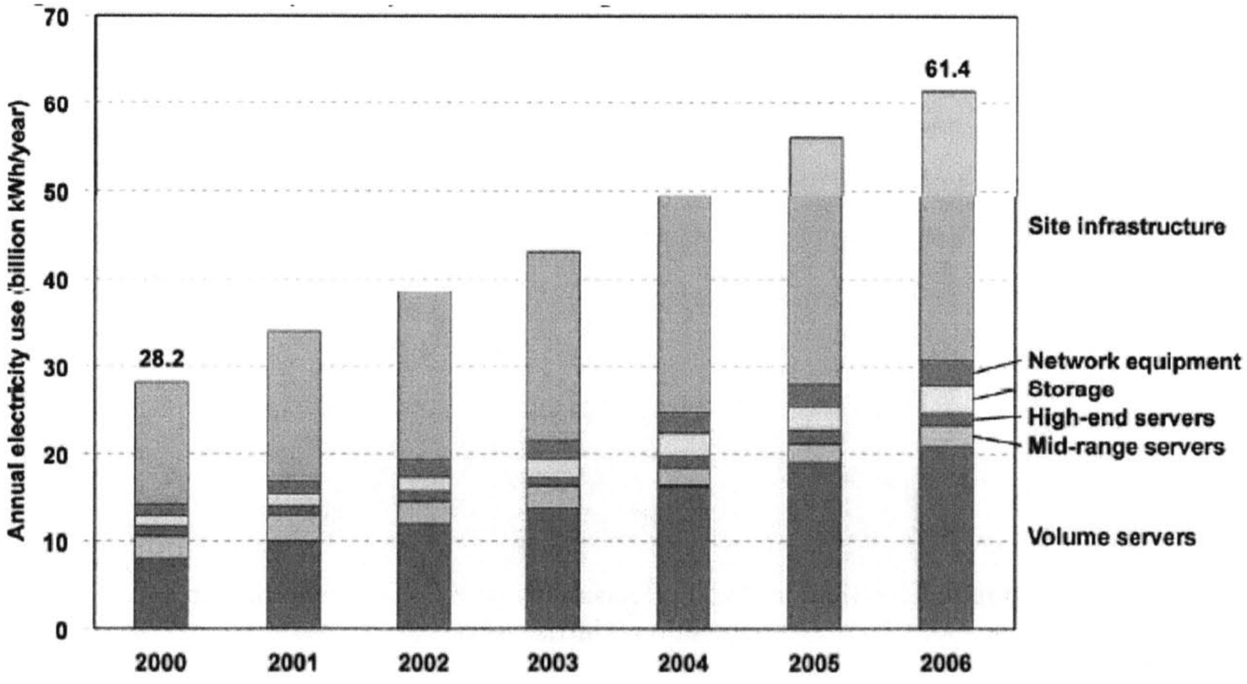


Figure 7: Electricity Use by End-Use Component (Fanara, et al., 2007)

Additionally, the IT load contributes to over 70% of the heat generation (Evans, 2004)

The two big problems we need to solve first, the Power Wall and the Memory Wall, are at the bottom of the computational chain

The Power Wall Problem: Moore's Law leads to unmanageable Power Density

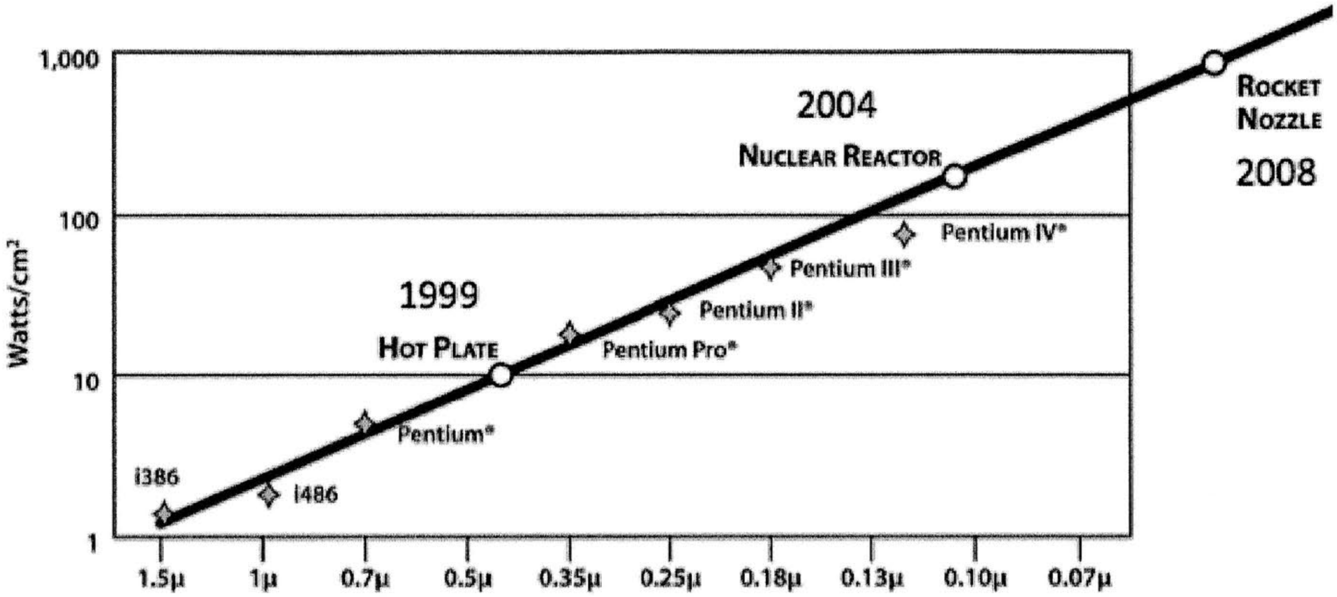


Figure 8: Power Density Becomes Too High to Cool Chips (Cong, 2008)

The Memory Wall Problem

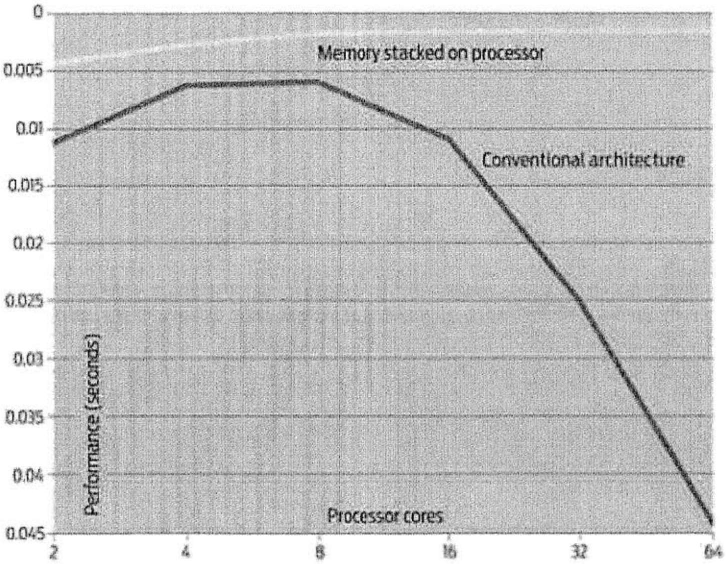


Figure 9: The Memory Wall affects many core CPUs due to resource contention (Harris, 2008)

Many HPC applications do not get linear benefits from multi-core processors.

> *"First of all, as chip geometries shrink and clock frequencies rise, the transistor leakage current increases, leading to excess power consumption and heat... Secondly, the advantages of higher clock speeds are in part negated by memory latency, since memory access times have not been able to keep pace with increasing clock frequencies. Third, for certain applications, traditional serial architectures are becoming less efficient as processors get faster (due to the so-called Von Neumann bottleneck), further undercutting any gains that frequency increases might otherwise buy. In addition, partly due to limitations in the means of producing inductance within solid state devices, resistance-capacitance (RC) delays in signal transmission are growing as feature sizes shrink, imposing an additional bottleneck that frequency increases don't address." (Borkar, et al., 2005)*

Contributions: How do we get a small PUE?

Our task is to reduce overall power use in the Datacenter and make the computation and power use we are left with more efficient. To get to where we want to be we need to follow some simple guidelines. We need to perform a bottom-up analysis and make improvements all along the power chain. The path to energy-efficiency at the HPC Datacenter can be found by following these 10 simple rules which we can apply across our HPC subsystems and cascade up to eliminate power use and improve our PUE:

1. The operation not performed is the most energy-efficient.
2. Implement HPC equipment maintainer / user behavior modifications.
3. Categorize and maximize things that lend themselves to consolidation and distribution to leverage hybrid architectures.
4. Give the big problems their due emphasis but also solve the lots of little problems.
5. Find a way to effectively utilize idle cycles for computation.
6. Compile code locally to maximize resource usage.
7. Use the most numerically efficient approach.
8. Convert to DC once and stay there.
9. Uncouple PE to RAM consolidation to ameliorate the Memory Wall.
10. Overlap and integrate computation and communications.

The path to a solution

The Datacenter should be viewed as an integrated facility with interdependencies among the 4 major subsystems (Power, Infrastructure including HVAC, IT Hardware, and Software / User

Behaviors). Improvements, even minor ones, in one area could have profound impacts on another subsystem. A bottom-up analysis will lead us to an effective way to determine which design decisions will trigger radical changes in another subsystem. For instance, in our Figure below, the triangle encloses the areas which we need to address in the Green HPC Center; each layer up the pyramid benefits from the savings below it and can then be further optimized to the lower power load. A minor 1% improvement in software efficiency will cascade up our chain until it represents a huge savings at our PDU point. This is the most inflexible piece of equipment to swap out or modify after it is in place so we need to determine what rating we need there through this analysis.

For instance, after our reverse-chain analysis, if we discover that we have saved a MW on the power input side that may result in the removal of a chiller system which will substantially reduce our CAPEX and if our PUE is close to 1, it will remove another MW of power use. Removing an HVAC system is not a very flexible option but it would have a great impact on our power use. In the same way, changing a software routine so it is more computationally efficient is easy to do after the system is in place, and it could have considerable impact on our power use as well.

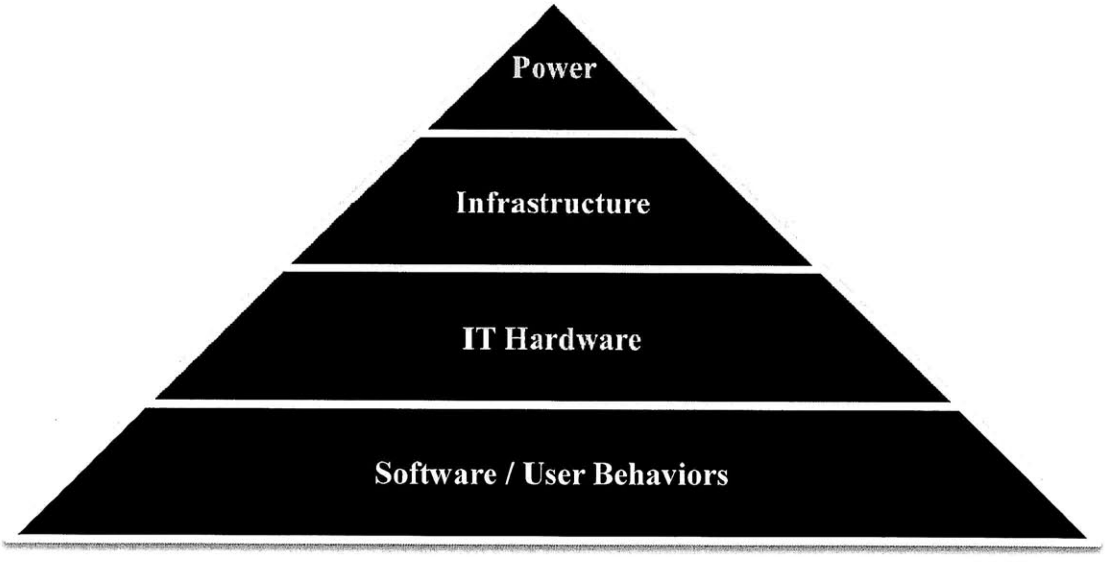
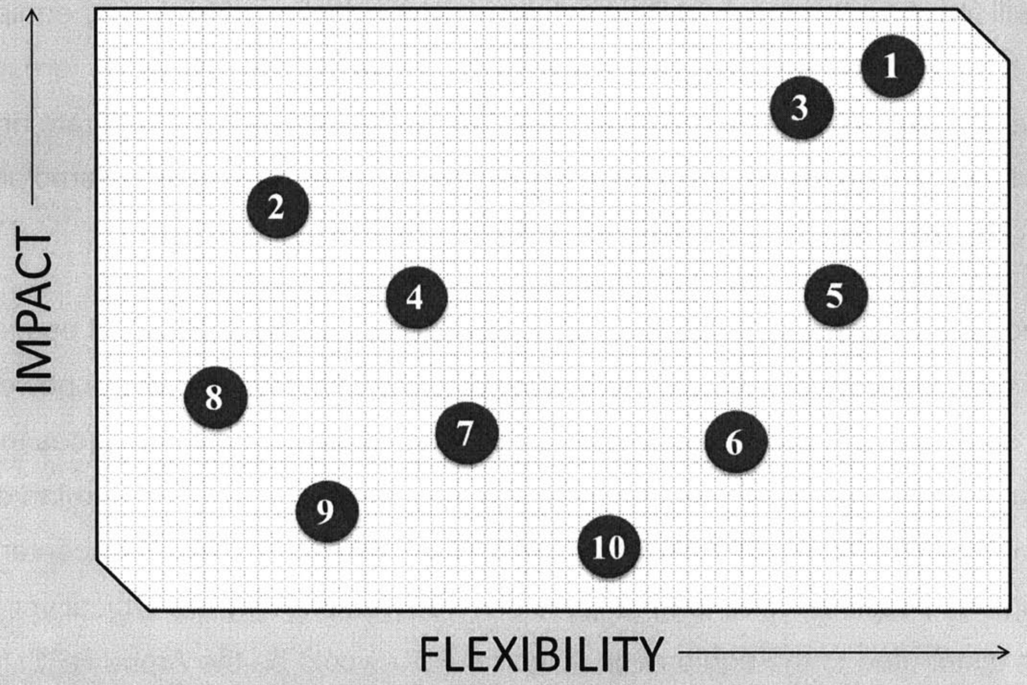


Figure 10 The HPC Datacenter pyramid, showing the interdependencies of the 4 major subsystems.

10 Rules for Energy Efficiency



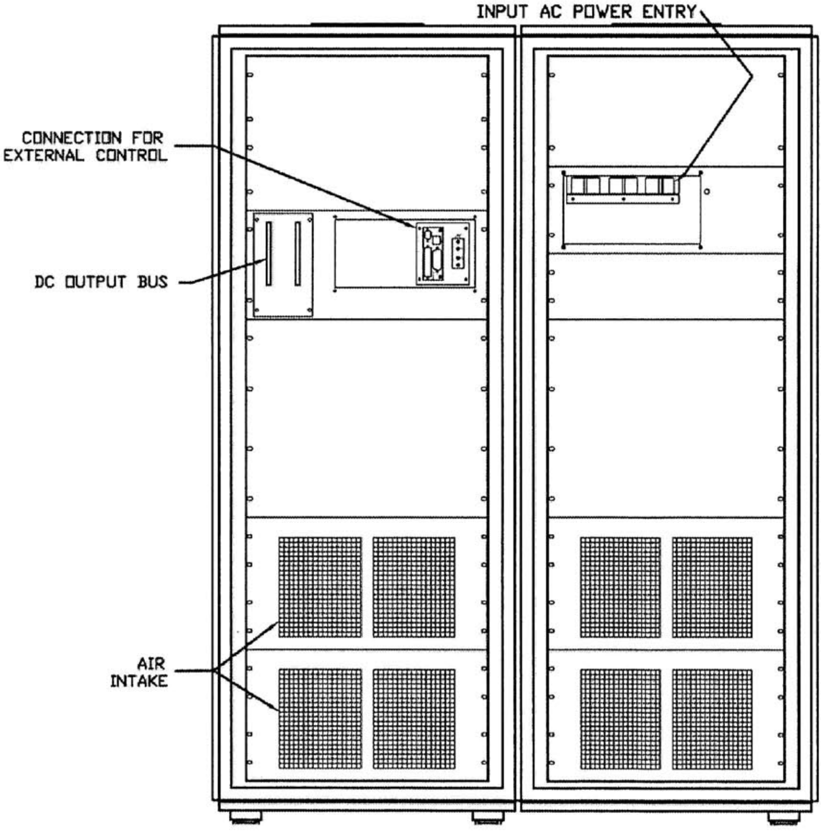
1. **The operation not performed is the most energy-efficient.**
   This could be something as simple as removing the requirement of ECC on your memory modules.
2. **Implement HPC equipment maintainer / user behavior modifications.**
   User HPC equipment maintainer behavior modifications. Educate those responsible for equipment maintenance.
3. **Categorize and maximize things that lend themselves to consolidation and distribution to leverage hybrid architectures.**
   Categorize and maximize things that lend themselves to consolidation and distribution and leverage hybrid architectures.
4. **Give the big problems their due emphasis but also solve the lots of little problems.**
   Give the big problems their due emphasis but also solve the lots of little problems. This is the only way to get to exascale.
5. **Find a way to effectively utilize idle cycles for computation.**
   Find a way to effectively utilize idle cycles for computation. You are paying for them anyway so you may as well use them. This can be achieved through established best practices use of virtualization, load balancing, and queuing systems.
6. **Compile code locally to maximize resource usage.**
   Use local compiling to maximize resource usage.
7. **Use the most numerically efficient approach.**
   This can be achieved by deciding to represent numbers in the target precision and accuracy throughout the processing cycle. You do not need to use 10 decimal places if your final answer need only be delivered to 2 decimal places.
8. **Convert to DC once and stay there.**
   Energy savings are best realized when AC/DC conversion is negated.
9. **Uncouple PE to RAM consolidation to ameliorate the Memory Wall.**
   Uncouple PE to RAM consolidation to ameliorate the Von Neumann bottleneck.
10. **Overlap and integrate computation and communications.**
    Prevent a data-starved or network-starved imbalance.

Chap. 2: HPC Hardware trends and recommendations for increased efficiency

In this section we will not describe the myriad options available to the system designer. Instead we will focus on the most realistic options, and a selection of tools and equipment classes which the HPC system designer will likely have available to him in the foreseeable future. We will then focus in more detail on areas where there is room to innovate and improve efficiency. Where there is a standard practice, we will describe an exemplar system.

PDUs, plug-strips, and server-level UPS

A good Datacenter design will match the power provisioning to the anticipated usage ("rightsizing" with a little margin built in for expansion). Starting from where the power enters the building, some centers use building-level UPS. From there the electricity goes to a power distribution unit which then breaks out the power to multiple electrical panels to fan out and simplify delivery of electricity to the IT racks. Generally speaking, a good Datacenter keeps its voltage as high as possible for as long as possible. 480V and 13.2kv are typical voltages before it gets distributed to the equipment. A typical PDU would be the Amrel HPS 150KW. It takes 480v 3-phase AC in and then breaks it out to lower voltage AC and DC rails.



CONNECTOR EXT. PRG.

TERM	PARAMETER	TERM	PARAMETER
1	REF GND	20	REF GND
2	REF GND	21	REF
3	VREF EXT	22	·REF EXT
4	TVREF EXT	23	T·REF EXT
5	VO2	24	·O2
6	REF CAL	25	NC
7	GND	26	+5
8	POWER	27	PGM L·NE
9	THERMAL	28	STANDBY
10	·NTERLOCK	29	PHASE LOSS
11	CUR CTL	30	VOLT CTL
12	STANDBY/ALM	31	RESERVE
13	ALM	32	OCT
14	EXT CTL	33	·NT CTL
15	FUSE	34	OVT
16	RESERVE	35	RESERVE
17	START	36	RESERVE
18	CLEAR	37	·NTERLOCK SET
19	STOP		

CONNECTOR JS2

TERM	PARAMETER
1	VO1REM-
2	VO1REM+

Figure 11 Rear view of Power Distribution Unit

Startup or boot time is not a significant percentage of the overall time of computer use anymore so turning machines on and off remotely has become an accepted practice in Datacenters. IP addressable plug strips, WOL (Wake-on-LAN) command scripting, or similar devices are used for turning equipment on and off.

UPS: Improving power usage

A certain percentage of the computational equipment is connected to a UPS (an uninterruptible power supply). The mission-critical servers will be put on this circuit. The way a UPS works is that in the case of a power loss it will allow you to keep your machinery running until such point you can do a "safe shutdown" or backup power generating systems can turn on prior to when the UPS batteries are depleted. This was necessary in the days when systems would reboot back into an unusable state after a hard power outage. The building-level UPS is substantially no longer needed, because if in the extremely unlikely loss of power to the IT equipment, you will only lose data up to your last checkpoint. This is a tolerable risk considering the risk / reward tradeoff and the cost of equipment. The server-level UPS equally allows you a certain period of time in which you can operate your machine and perform a "safe" automated shutdown or turn on your on-site generation equipment, usually in the form of diesel generators. Use of the UPS datacenter-wide is an anachronistic convention. They certainly aren't currently worth the cost of the electricity associated with converting AC to DC, storing it in batteries, and then converting it back to AC in order to power the servers anymore.

Network infrastructure

This includes switches and routers and to a lesser extent firewalls and hubs. A typical network switch would be a Broadcom E – series switch (seen in Fig. 14). This switch incorporates the IEEE energy-efficient ethernet standard IEEE 802.3az.

Trends

We anticipate density and consolidation to continue on its substantially Moore's Law following trajectory. The network-centric version of this Law tends to be of a coarser grain size. Commodity network speeds increased from 10Mbs/ to 100Mb/s in 5 years and then from 100Mb/s to 1Gb/s in just 3 years recently. The next standard that will be in general use once the price comes down in the next few years will be 10Gb/s Ethernet (IEEE Std. 802.3ae-2002). Like GigE Ethernet, 10GE will be offered in both optical and "twisted-pair" formats. It is expected that one will be able to upgrade their GE chassis to 10GE with a card swap.

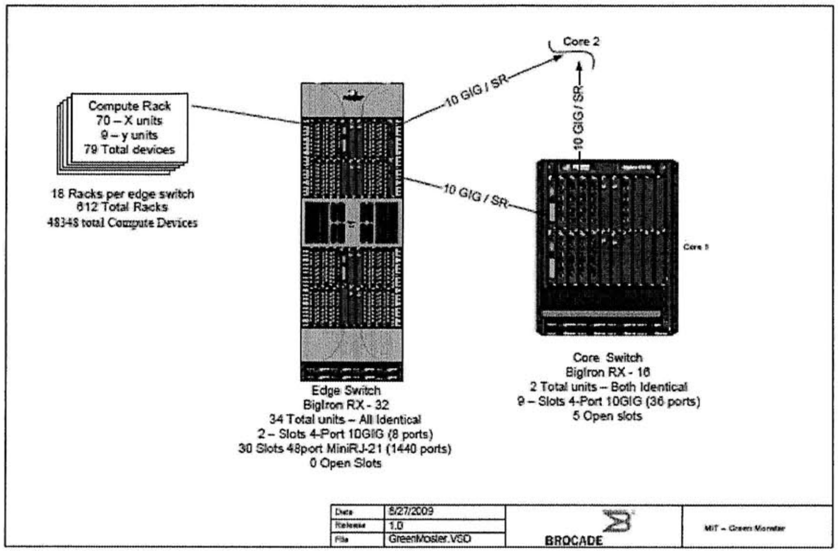
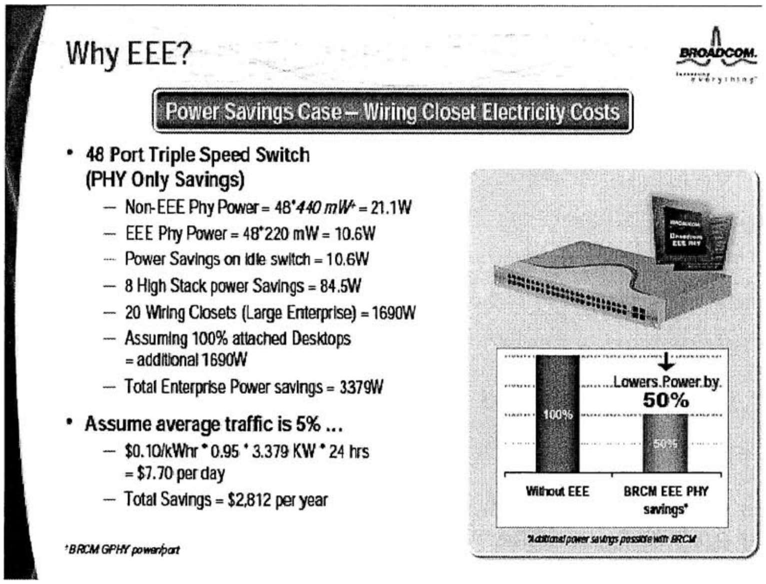


Figure 12 A typical multi-level (Core and Edge) switch Datacenter layout

Recommendations

The path to follow is to adopt the new EEE standard switches. They achieve the dual goals of getting better resource utilization and lower power use by as much as 50% for certain applications.



Why EEE?

Power Savings Case – Wiring Closet Electricity Costs

- **48 Port Triple Speed Switch (PHY Only Savings)**
  - Non-EEE Phy Power = 48*440 mW* = 21.1W
  - EEE Phy Power = 48*220 mW = 10.6W
  - Power Savings on idle switch = 10.6W
  - 8 High Stack power Savings = 84.5W
  - 20 Wiring Closets (Large Enterprise) = 1690W
  - Assuming 100% attached Desktops = additional 1690W
  - Total Enterprise Power savings = 3379W
- **Assume average traffic is 5% ...**
  - $0.10/kWhr * 0.95 * 3.379 KW * 24 hrs = $7.70 per day
  - Total Savings = $2,812 per year

*BRCM GPHY powerpart

Figure 13: EEE standard switches save power (Broadcom, 2001)

Network interconnect

Trends

Almost 90% of the Top500 list uses one of two interconnect standards; gigabit ethernet (a twisted pair copper IEEE 802.3ab standard) and Infiniband (an industry-driven optical standard). Gigabit ethernet has effective bandwidths of approximately 800 Mb per second and latencies around 12 μs. Infiniband costs considerably more than gigabit ethernet but has better bandwidth and lower latency. A typical 4x SDR Infiniband switch will run at 8Mb/s and have latencies in the order of 200 nanoseconds while costing at least 10 times an equivalently sized GigE switch. Optical technologies, in general, have better internal efficiencies than the copper equivalent. (Top500.org, 2010)

Recommendations

Both of these technologies can benefit through better management of data communications. The interconnect is used in HPC to communicate to other computation nodes as well as to the network attached storage (NAS), or to the storage area network (SAN). So even minor improvements in communication efficiencies would have a broad positive systemic effect. These can be achieved through improvements on device driver coding. A promising candidate for a low-latency GigE device driver is the GAMMA protocol that can replace traditional TCP communications in many areas.

Compute subsystems

Anything that improves computational efficiency will improve energy efficiency. We will describe a number of approaches that will get us to an energy-efficient and computationally powerful network solution.

Component	Peak Power (watts)
CPU	80
Memory	36
Local Disks	12
Peripheral slots	50
Motherboard fan	25
PSU losses	10
	38
TOTAL	**251**

A typical server power breakdown is listed above. There are many places we can improve energy efficiency in this model.

Trends

This is where a considerable amount of the energy efficiency improvements can take place. Modern x86 and x86–64 architectures leak current because of the way they are designed.

This problem has been compounded in recent years as we have incorporated ever-increasing amounts of transistor density. Over 450 of the Top500 use these architectures.

Recommendations

Investigate alternate CPU architectures. The most energy-efficient systems on the Green500 use non-x86 CPUs. The IBM Blue Gene, STI Cell Broadband Engine, and GPU centric systems dominate the Green500 list. (Green500.org, 2010)

Server PSUs

Trends

Most commodity switching PSUs allow you to select from 120v or 240v AC. They run at around 65 - 85% efficiencies and add a point of electrical / mechanical failure into the system.

Recommendations

Purchase the most efficient switching power supplies, often those indicated with an Energy Star certification. Generally, the most efficient conversion for these is to take power in at 240v AC and then downconvert to the various voltages required on the server motherboard.

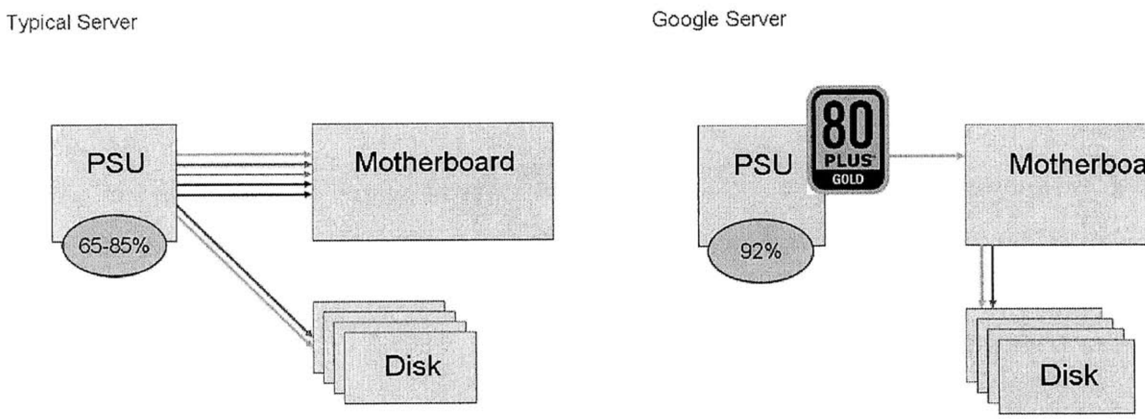


Figure 14 and Figure 15 Comparison of Google server vs. typical server

Google innovated on this by converting the power to DC once and then driving the disks off of the motherboard. (Google, 2009)

Memory subsystems

Trends

There are a number of interesting memory technologies on the far horizon. None of these technologies are predicted to make their way into high-performance computing this decade though. Instead we are destined to work within the memory allocation and interaction paradigm with which we have become quite accustomed. This is the predictable progression of memory on enterprise motherboards to follow a Moore's Law like path. Indeed, Moore's Law was originally invoked to describe memory transistor counts rather than CPUs. To this end, we can expect memory modules to double in density approximately every 18 months. This won't keep up with the Memory Wall disparity developing further though. That has more to do with matching speeds and bus widths. From 1986 to 2000, CPU speed improved at an annual rate of 55% while memory speed only improved at 10%. That trend will continue for some time.

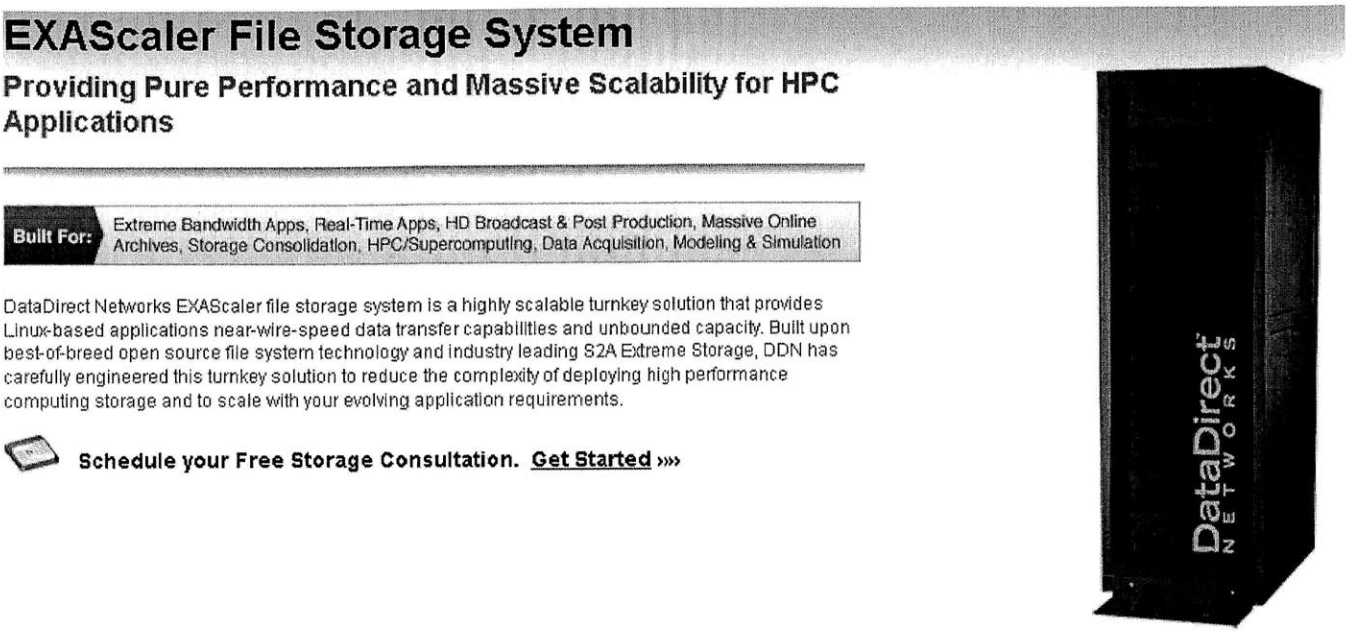
Recommendations

Rather than follow a path that has a predictably troublesome collision with the Memory Wall, designers should rethink their processor to memory ratio. For instance, for memory-intensive operations, consider 8 core or less CPUs to ensure each processing core has sufficient bus bandwidth to memory. This is a strategy successfully employed on the IBM Blue Gene systems and contributes significantly to their placement in the top of the Green500 List.

Storage subsystems

Trends

Almost half of the Top500 list used DataDirect Networks as their Infiniband SAN provider. This is significant since it points to a trend towards inexpensive Infiniband becoming a de facto standard at HPC centers. An exemplar here would be the DDN Exascaler. (Data Direct Networks, 2011)



EXAScaler File Storage System

Providing Pure Performance and Massive Scalability for HPC Applications

Built For: Extreme Bandwidth Apps, Real-Time Apps, HD Broadcast & Post Production, Massive Online Archives, Storage Consolidation, HPC/Supercomputing, Data Acquisition, Modeling & Simulation

DataDirect Networks EXAScaler file storage system is a highly scalable turnkey solution that provides Linux-based applications near-wire-speed data transfer capabilities and unbounded capacity. Built upon best-of-breed open source file system technology and industry leading S2A Extreme Storage, DDN has carefully engineered this turnkey solution to reduce the complexity of deploying high performance computing storage and to scale with your evolving application requirements.

Schedule your Free Storage Consultation. Get Started »»

Figure 16 *A Contemporary Scalable HPC Storage System*

Tape

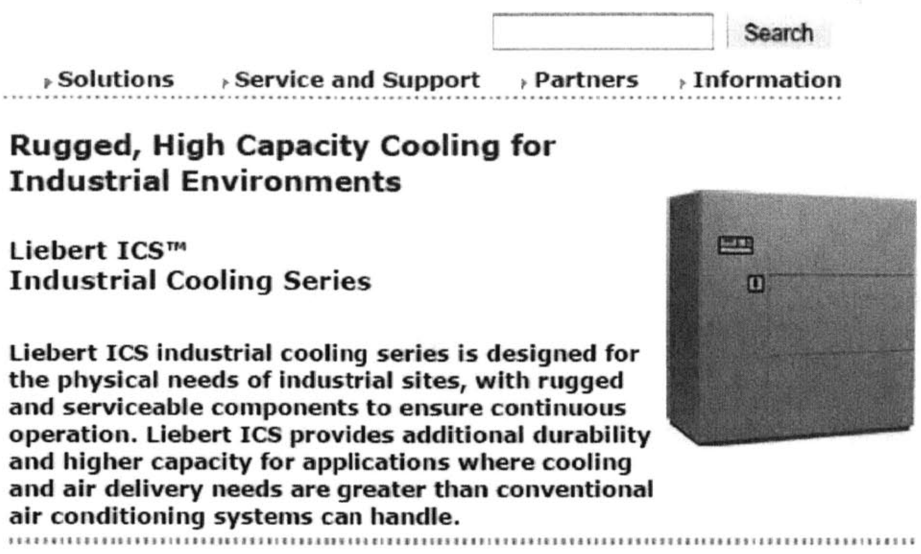
Tape systems have historically been used as an archive for what the Datacenters have on disk, but a surprising number of centers are starting to use it as a space, energy, and cost efficient way to store data outside of disk usage. The latency is quite high but that appears to be the only major tradeoff.

Cooling subsystems

This is where a considerable amount of work can be done in improving energy efficiency in the Datacenter through an addition-by-subtraction strategy. If we find a way to generate less heat in the compute subsystem, then we would have less heat to deal with on the cooling subsystem. Less quantity (usually measured in cubic feet per minute or CFM) will make our heat removal job exponentially easier.

Trends

Most existing Datacenters are air-cooled but trends in HPC are substantially towards water-cooling systems, in particular, in-rack water-cooling systems (server-side economizers).



Rugged, High Capacity Cooling for Industrial Environments

Liebert ICS™
Industrial Cooling Series

Liebert ICS industrial cooling series is designed for the physical needs of industrial sites, with rugged and serviceable components to ensure continuous operation. Liebert ICS provides additional durability and higher capacity for applications where cooling and air delivery needs are greater than conventional air conditioning systems can handle.

Figure 17 Contemporary High Capacity Cooling Equipment

Recommendations

While liquid cooling is more efficient for large centers, air-cooling is sufficient for smaller ones, or ones with a considerably lower amount of heat to remove. This is the payoff with energy-efficiency improvements on the IT equipment; you will be able to buy considerably cheaper air-cooling systems if you can get the "waste" heat down to a level that is manageable with such systems. This will make hardware management easier as well.

Server Room layout

Most existing Datacenters are air-cooled. There is considerable standardization and agreement in this area. All Datacenter racks are 19" racks so floor tiles have been designed to accommodate this layout. Virtually all centers use perforated floor & ceiling tiles over a raised floor. This affords the use of down-flow CRACs or CRAHs and a Cooling Economizer. Some centers use in-rack water-cooling. Convention dictates that an air-cooled Datacenter will have rows of server racks with their air intakes and exhausts oriented in an alternating pattern to ensure that the cold and hot air masses don't mix. Conventional practice guidelines are to automate or proactively manage the power settings of this equipment based on humidity and temperature readings.

The Hot Aisle Cold Aisle Model

The Hot Aisle Cold Aisle Model is defined by the design and coordination of its HVAC systems. It's success or failure of implementation relies on its emphasis on minimizing obstructions to proper airflow. Blanking panels, often unused by operators and designers are key to ensuring proper airflow.

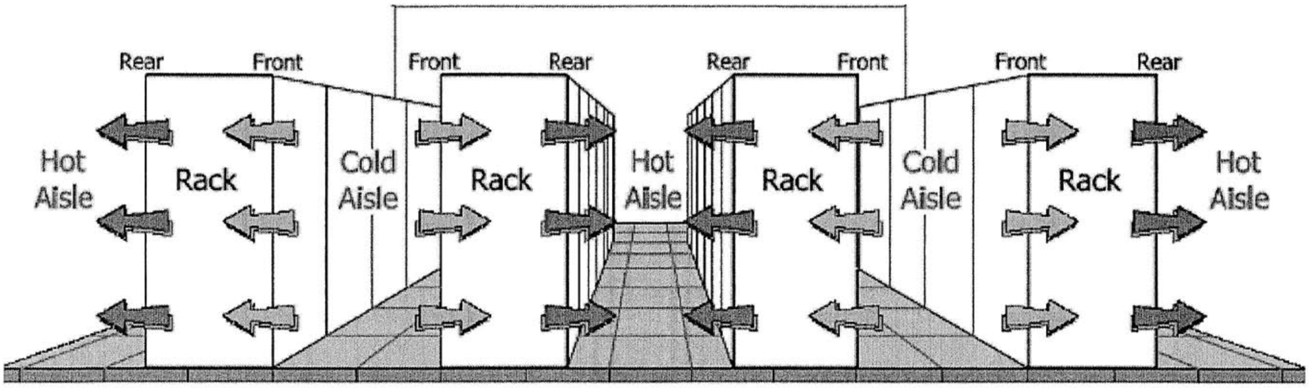


Figure 18 Typical Datacenter HVAC Hot Aisle / Cold Aisle Layout
(Fanara, et al., 2007)

Miscellaneous equipment energy efficiency

Most modern centers are utilizing a "black-out" procedure where the lights are off for the majority of time in the Datacenter.

User behavior modifications

The majority of these energy efficiency improvement recommendations can be implemented either before the equipment is emplaced or as part of the operation by the people who manage and program the same. This relieves us of the requirement to educate a large number of end-users up on techniques they may not be familiar with or may be reluctant to learn. You have considerable control over the low-level operation of the equipment in an HPC center.

Summary

Utilizing the recommended approaches in making your hardware more efficient will pay dividends in the following 10 Rules For Energy Efficiency contributions areas; 1 - 5, and 8 - 10. In particular, Rule 1, which states that the work not performed is the most efficient, is a maxim that urges designers to reduce their power and cooling footprint to a point that they can remove equipment from their design at the more inflexible level of the design pyramid and thereby have a substantial impact of TCO.

Chap. 3: HPC Software trends and recommendations for increased efficiency

In this section we will identify areas of improvement in software design and practice as well as suggest operational method improvements.

Numerical representations improvements

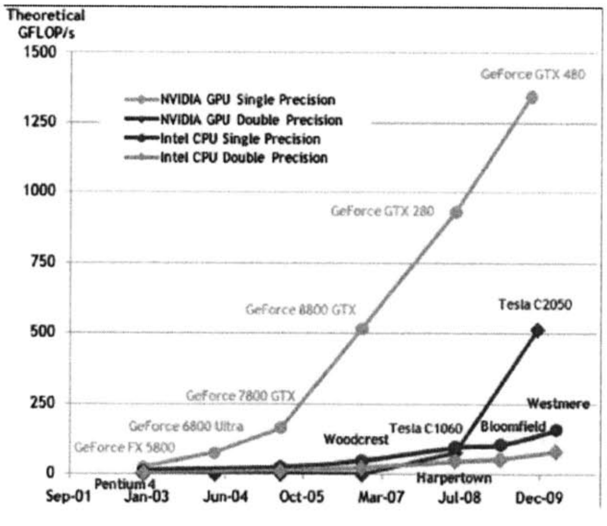
Trends

Designers are purchasing equipment capable of more native double precision work. This is demonstrated in the proliferation of multicore systems in the Top500. (Top500.org, 2010)

Recommendations

The CPU (especially the x86 variety) is the single biggest energy hog in the chain. Keep your calculations in single float precision (or better yet half) where you can to better utilize the hardware. Target your single precision or half precision capable registers for the bulk of your computation and then fix your precision by using iterative refinement with fewer double precision registers.

GPU operations are always faster at single precision

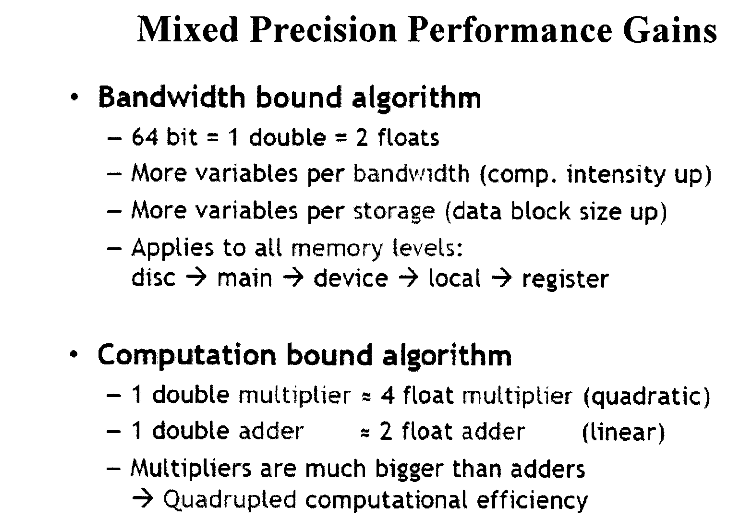


Cuda Programming Guide

Figure 19: Single precision vs. double precision operations

Mike Clark's work at Harvard is typical. (Clark, La Plante, & Greenhill, 2011)

If you need double precision, which many applications don't, you can fix it by using a technique called iterative refinement pioneered by Goddeke. It turns out this is not only faster but demonstrably more accurate.



Mixed Precision Performance Gains

- **Bandwidth bound algorithm**
  - 64 bit = 1 double = 2 floats
  - More variables per bandwidth (comp. intensity up)
  - More variables per storage (data block size up)
  - Applies to all memory levels:
    disc → main → device → local → register
- **Computation bound algorithm**
  - 1 double multiplier ≈ 4 float multiplier (quadratic)
  - 1 double adder ≈ 2 float adder (linear)
  - Multipliers are much bigger than adders
    → Quadrupled computational efficiency

Figure 20: Performance Gains through precision change adoption
(Strzodka & Goddeke, 2008)

Use of Graphics and Floating Point Extensions

Prof. Hank Dietz at the University of Kentucky has long contributed to the algorithmic improvements that you can get with Intel and AMD CPUs (x86 and x86-64). He pioneered a technique called SWAR (SIMD within a register) to get parallelism within a CPU which complemented the parallelism we were already getting across the cluster. He recently pioneered a technique called MIMD on GPU (MOG), which again takes an existing architecture and through algorithmic processes makes better use of the hardware by getting more calculations out of a cycle. If you have to power all parts of these chips, like the SSE2 registers, then you may as well maximize their utility. (Dietz H. , 2000) (Dietz H. G., 2010)

Recommendations

Use the specialized headers associated with your particular hardware and recompile your applications to take advantage of the resources this makes available. HPC centers have employees who are there to make particular pieces of application code compliant with their operating systems and middleware so the expertise is there to maximize these optimizations.

System tuning

There are a number of things that you can turn on in the BIOS, including ACPI and AHCI for the hard drive controller. (LessWatts.org, 2011)

- Enable the power aware SMP scheduler
- Use SATA link power management

- Turn off or remove WiFi, Bluetooth and other unused chipsets
- Set the appropriate "Wake-on" settings
- The BIOS can be updated by a script that writes these settings to the CMOS from the Linux command prompt.

Operating systems

Trends

Because Linux represents the operating system on over 95% of the Top500 list supercomputing sites (Top500.org, 2010) we will limit our discussion to energy efficiency improvements to Linux. While there are some operating systems that are more energy efficient than Linux they cannot perform some of the basic functions we need in HPC and would therefore not be an appropriate part of our solution. Considering the number of energy efficient applications libraries and middleware that Linux supports, this exclusion will prove to be an appropriate selection criterion. Using the latest even-numbered kernel will glean you a Watt or 2 back per machine. Linux has a couple of features and utilities, notably *tickless* and *powertop*, that affords some energy efficiency improvements and facilitates some tool uses that help us identify inefficient parts of our code or operating techniques. As we have said, most HPC computers run at above 90% utilization so some of these tools (like virtualization) won't be of use to us.

Recommendations

Linux improvements are predictable and rapid. There will be consistent upgrades to performance and energy usage reductions at every iteration and release of new versions of the system. EE tuning can be achieved through "turning on" a few things in the kernel or patches that are specifically designed to enable a certain functionality.

Queuing systems, Schedulers and Control Systems

Because of the aforementioned voltage and current leak problems in modern CPUs, there is only so many ways we can make the CPU more efficient. This means it is of increasing importance that we keep the CPU busy (at 99% usage) wherever possible. We need to minimize context switching since cycles wasted on things like "garbage collection" and data error checking don't translate into getting real work done.

Trends

Queuing systems, processor affinity, and application code improvements are getting better at keeping the CPU busy. This will improve with subsequent revisions of these packages.

Recommendations

SGE (Sun Grid Engine, originally developed by Sun Microsystems) is one of the most popular queuing systems on Linux. It is sufficiently representative of all queuing systems for us to be able to confirm we can use these in an energy-efficient manner. A properly installed queuing system or batch scheduler will associate the appropriate "nice" level with job to allocate priority correctly and will shut down nodes not in use via their power-saving utilities, mostly through a wake-on-LAN functionality. (Dolz, Fernandez, Mayo, & Quintana-Orti, 2010)

Even an efficient server still consumes half of its full power when idle

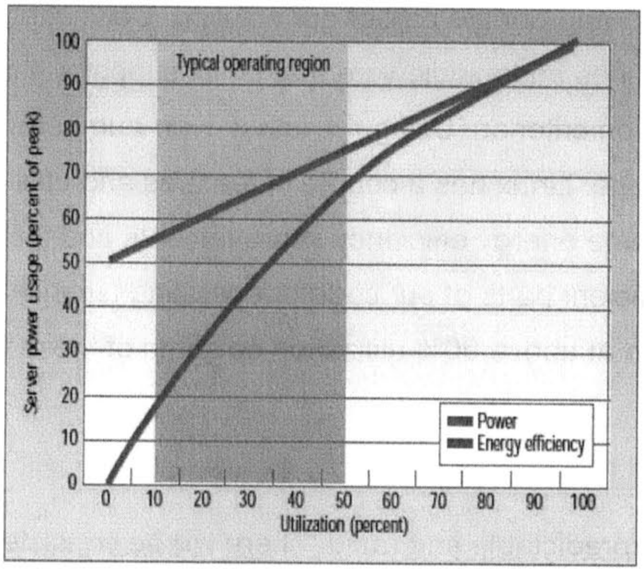


Figure 21: Server power usage vs. utilization (Barroso & Hölzl, 2007)

Application Code

This is probably the most subjective area that we need to address. A proper high-performance code needs to be written in such a way that it makes the most calculations per CPU cycle and the I/O portion needs to be written so that data reading or writing out to disk takes approximately the same time as the calculating portion in order for you to make the greatest use of the CPU (no idle cycles). This can be achieved through the use of code profiling and autotuning tools like PyCUDA or similar. A properly written application program will overlap the communication time and the computation time by processing the appropriate amount of data to achieve this. This will mask any latency associated with the processing time by removing any "data starvation" seen at the CPU.

Energy monitoring systems

Most Linux systems come with some DVFS-aware package these days that allow you to either report inefficiencies or react to the condition according to some rules you put on the system. For instance, Ganglia and Nagios are useful tools to easily visualize the power usage of the various subsystems in your cluster and will allow you to tune parameters appropriately.

Virtualization

Trends

There is a significant movement in the Datacenter and data provisioning business towards virtualizing hardware through an abstraction layer (or a "hypervisor") to reduce power used during idle time. This is usually achieved by applying some middleware into your operating system stack that manages clock speeds and resource allocation. Popular packages for this under Linux are VMWare tools, Mosix, and Virtualbox. These tools give the operator some automation of resource allocation on a given server that will reduce power consumption in low usage periods. This makes sense in a web services provider Datacenter because of the inconsistent nature of their computational loads. However, in an HPC Datacenter, CPUs are normally running at 99% utilization consistently. It can therefore be considered harmful to performance to add an abstraction layer which could be deemed unnecessary.

Recommendations

Our use of virtualization comes in the form of a much more coarse-grained approach. A given server is either on or off, but if it is on, it is being used at its full capacity. We are achieving this through power system management and are not as concerned with the more detailed management that can be achieved through the use of virtualization tools that come with a broad, indiscriminate, performance penalty.

Summary

Utilizing the recommended approaches in making your software more efficient will pay dividends in the following 10 Rules For Energy Efficiency contributions; 1 - 7, and 10. In particular, tuning your application code so that it takes advantage of your PE's native precision capabilities and overlaps communication and computation to make full use of the CPU cycles will yield significant raw performance and performance per watt benefits.

Maintainers and constructors purchase behaviors

The designers of HPC Datacenters generally plan the opening date of their facility to coincide with the ship date of the latest IT product from their vendor partner. This is usually around the time of a product announcement at the IEEE Supercomputing conference in November of every year. If a Datacenter can manipulate its construction schedule accordingly, they will have a month or so of "burn-in" time in advance of the conference so they can publish their LINPACK results on the new equipment to coincide with consideration for the Top500 list, announced at that conference. This period is generally run in some degree of secrecy since the vendor does not want officially published benchmarks prior to the product being available to the broader marketplace.

The reasons for this approach are now well understood. Moore's Law, and the resultant vendor roadmaps, dictate that a designer will have the state-of-the-art CPU, RAM, disk, and network interconnects available to them on the day of the product shipping and then the price curve will slowly erode until the next generation product is released, at which point the price will be considerably less when it is no longer the latest gear. A Moore's Law generation used to be 2 years but lately (the last 7 years) it has been 18 months, and some of the subsystems evolve at a slightly faster pace than that. We anticipate this trend to continue until no later than 2020, which is when the Power Wall and Memory Wall problems will be manifested to a state where we can no longer miniaturize our way to a functional solution at anything less than an untenable VAREX cost.

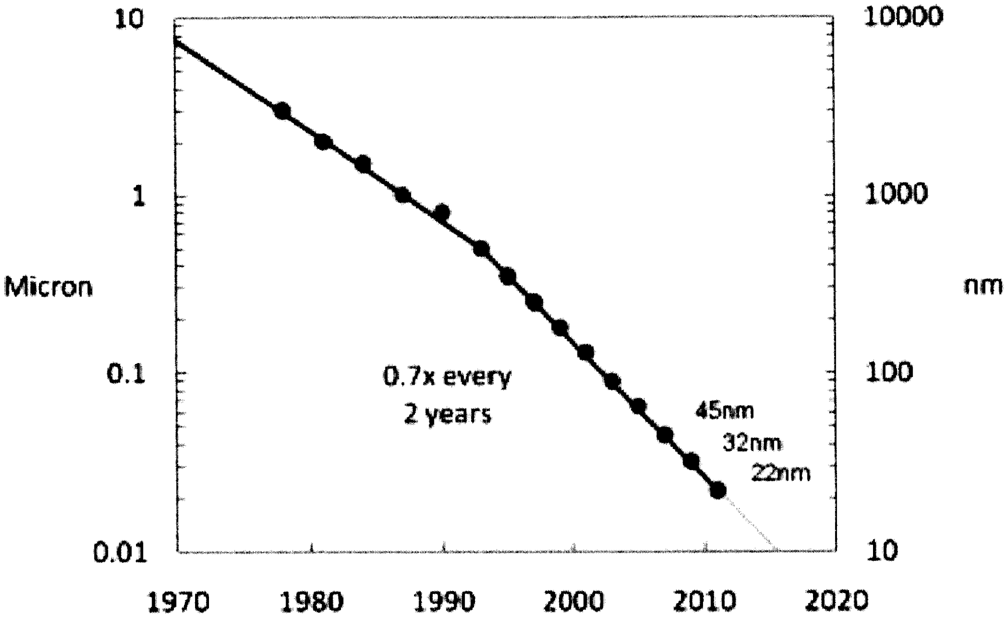


Figure 22: Reduction in scale follows Moore's Law (Intel, 2010)

HPC Center designers also need to ensure they are looking at TCO during the design process. It is tempting to go with a "lowest bid" solution since that very prominently appears to be the cheapest. But very often that is not the case. When the Army High Performance Computing Research Center came up for renewal in 2006 and was slated to move from the University of Minnesota to another University, through a DOD solicitation process, a number of Universities passed on it since it was encumbered with a number of legacy support restrictions. There were certain systems in the existing Center that needed to be emplaced in the new Datacenter and they were of a type that required unusual power formats and were very inefficient as well. So even if the initial estimates are that a particular purchase path will be cost-effective, designers must ensure that they are calculating in the VAREX over at least a 3-year window of system use, relative to an estimate of costs for more modern equipment. This phenomenon was particularly pronounced in this incidence. Even though the hardware was essentially "free," it was so power-hungry that it wouldn't make financial sense to keep that equipment in operation. Additionally, Datacenter owners can't expense VAREX so the upfront low cost of the equipment wasn't as much of attraction as a higher CAPEX / lower VAREX solution. (Dept. of the Army, U. S. Army Materiel Command, RDECOM Acquisition Center, 2006)

User behaviors

Studies have shown that the consumers of HPC resources have predictable usage patterns that HPC support personnel can adapt and accommodate. For instance, in the MIT Lincoln Laboratories LLGrid example below, most users requested less than the maximum available number of resources for a relatively short time (relative to the capacity of the system) and so were accommodated with a tailored request and queuing system.

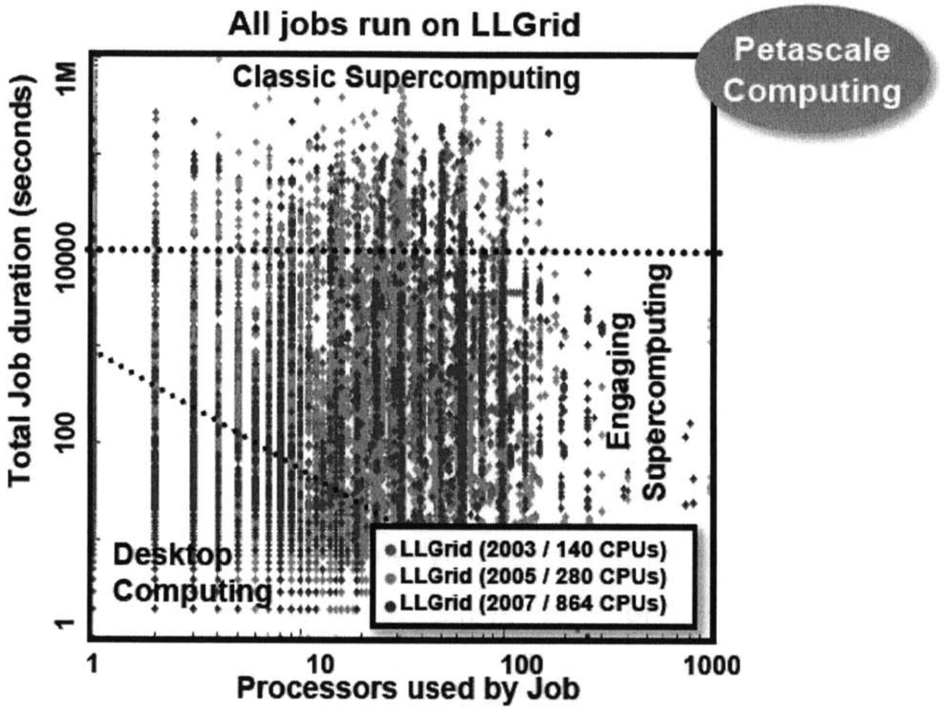
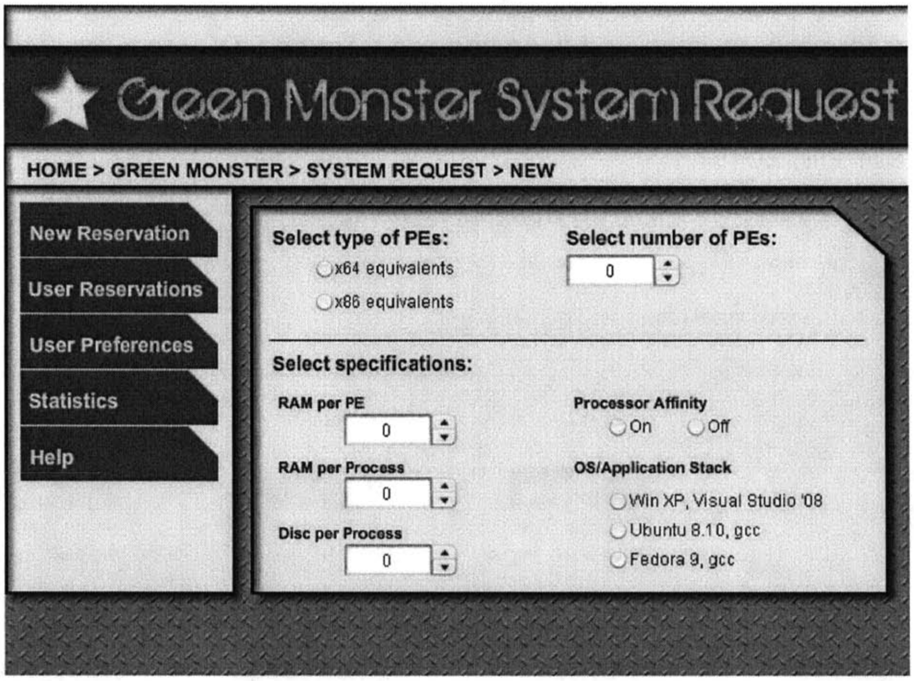


Figure 23 Historical job distributions on MIT LLGrid (courtesy J. Kepner).

The LLGrid used to suffer from resource allocation overkill, which they describe as inefficiencies from scaling bugs, time dilation, queue hacking, unmonitored jobs, allocation churn, and job overkill. If all resources are allocated to a "long job," we reduce the Unix nice level (priority) on a number of PEs to accommodate a short, interactive job. The LLGrid customer base predictably uses a small number of PEs for short periods of time while they are tuning their application code and then they use a large number of PEs for a large amount of time when they launch their production run. Lincoln Labs has accommodated this behavior through a process they term "Engaging Supercomputing". If resources are idle, a job requester can have them. If they are committed, you gain access to resources coincident with the length of the run. That way, there are no resources held in reserve waiting for a priority job to get launched. This would be wasteful of resources and power.

In a typical HPC scenario, a customer will compile his program on a single computer, perform a considerable amount of off-line development work, and then transfer it over to his account on the HPC servers. To get his production run, he will want to run his application simultaneously on as many processors as possible, usually utilizing a large input file and generating a large output file. An interface, usually implemented on a web page, is then accessed by the customer to request resources for their processing runs. It is appropriate to note that their request rarely gets denied; it may get queued or reduced in priority but given the historical usage patterns, one can anticipate that some level of service will be acceptable to the requester.


Green Monster System Request
HOME > GREEN MONSTER > SYSTEM REQUEST > NEW
New Reservation
User Reservations
User Preferences
Statistics
Help

Select type of PEs:
○x64 equivalents
○x86 equivalents

Select number of PEs: 0

Select specifications:

RAM per PE: 0

RAM per Process: 0

Disc per Process: 0

Processor Affinity: ○On ○Off

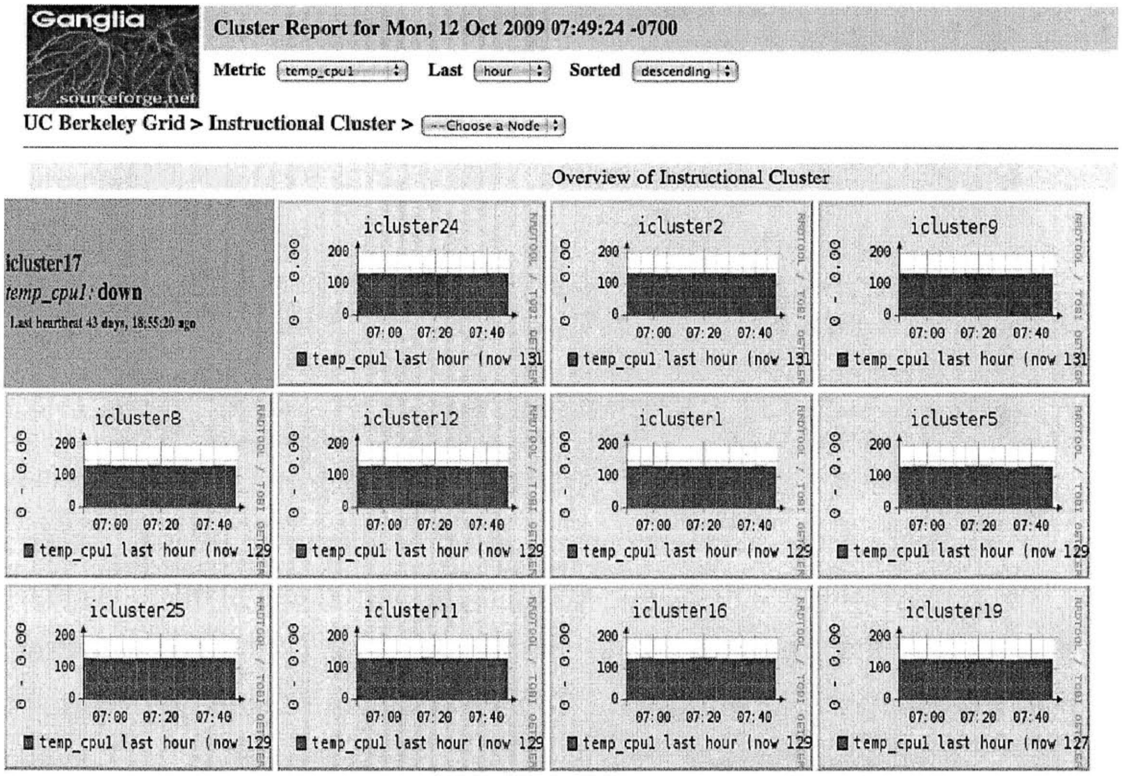
OS/Application Stack
○Win XP, Visual Studio '08
○Ubuntu 8.10, gcc
○Fedora 9, gcc

Figure 24 A typical user resource request page to submit HPC jobs

So that both users and systems administrators can monitor progress of the individual jobs and resources, a dashboard feature, typically implemented in Ganglia or Nagios or another publicly available product is used extensively. The administrators can perform certain functions from this interface if they see something that needs intervention, but substantially this tool is controlled through preemptive, proactive, and reactive settings and scripts.

This abstraction layer allows the System Administrators to offer PE equivalents to the users implicitly. As long as the users are getting their processor of choice, usually either 32-bit (in the form of an Intel x86 compliant processor) or 64-bit (in the form of an x86-64 compliant processor), their application code will run without issues. Additionally, alternate libraries and

APIs are offered to the user at this juncture, in case there was a particular software routine that was necessary for their code at this point. This can be the case when the programmer compiled his code against a non-standard library like a particular version of the BLAS library. The underlying request system scripts will make the appropriate links and include the path to the libraries so there will be no "run-time" errors reported to the system operator or user. This convention has been substantially made easier through the choice of Linux as an Operating System on both the user's development environment and the HPC shared resources. Furthermore, most scientific codes, in the broad areas of Computational Molecular Science, Cosmology, N-body simulations, Atmospheric Science, Bayesian Inference, Materials Science Modeling, and other research areas utilize standard math libraries that are available either publicly or through an accepted licensing program on HPC systems. Many of these, like BLAS and FFT, are encapsulated in one of the packages specified in the HPC Challenge suite. (HPC Challenge, 2010)



Cluster Report for Mon, 12 Oct 2009 07:49:24 -0700

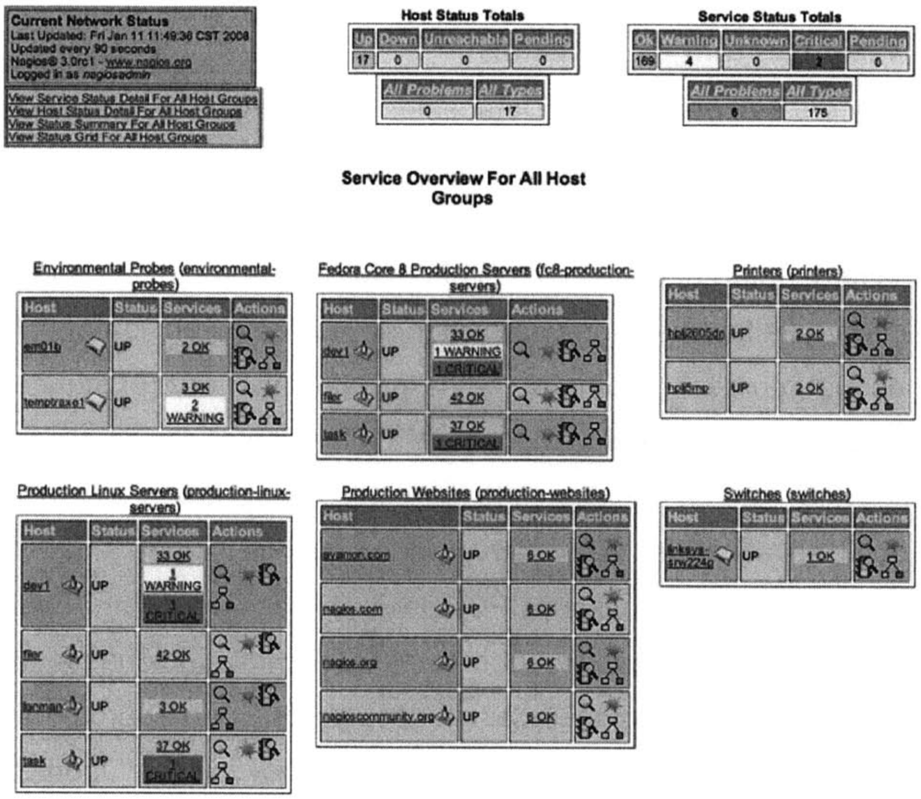
Metric temp_cpu1 Last hour Sorted descending

UC Berkeley Grid > Instructional Cluster > --Choose a Node

Overview of Instructional Cluster

icluster17
temp_cpu1: down
Last heartbeat 43 days, 18:55:20 ago

Figure 25 Users and Administrators monitor individual resource usage through a dashboard. They can select from a number of resource options and aggregate them over the cluster.



Current Network Status
Last Updated: Fri Jan 11 11:49:36 CST 2008
Updated every 90 seconds
Nagios® 3.0rc1 - www.nagios.org
Logged in as *nagiosadmin*

View Service Status Detail For All Host Groups
View Host Status Detail For All Host Groups
View Status Summary For All Host Groups
View Status Grid For All Host Groups

Host Status Totals

Up	Down	Unreachable	Pending
17	0	0	0

All Problems	All Types
0	17

Service Status Totals

Ok	Warning	Unknown	Critical	Pending
169	4	0	2	0

All Problems	All Types
6	175

Service Overview For All Host Groups

Environmental Probes (environmental-probes)

Host	Status	Services	Actions
em01b	UP	2 OK	
temptraxe1	UP	3 OK 2 WARNING	

Fedora Core 8 Production Servers (fc8-production-servers)

Host	Status	Services	Actions
dev1	UP	33 OK 1 WARNING 1 CRITICAL	
filer	UP	42 OK	
task	UP	37 OK 1 CRITICAL	

Printers (printers)

Host	Status	Services	Actions
hpl2605dn	UP	2 OK	
hpl5mp	UP	2 OK	

Production Linux Servers (production-linux-servers)

Host	Status	Services	Actions
dev1	UP	33 OK 1 WARNING 1 CRITICAL	
filer	UP	42 OK	
lanman	UP	3 OK	
task	UP	37 OK 1 CRITICAL	

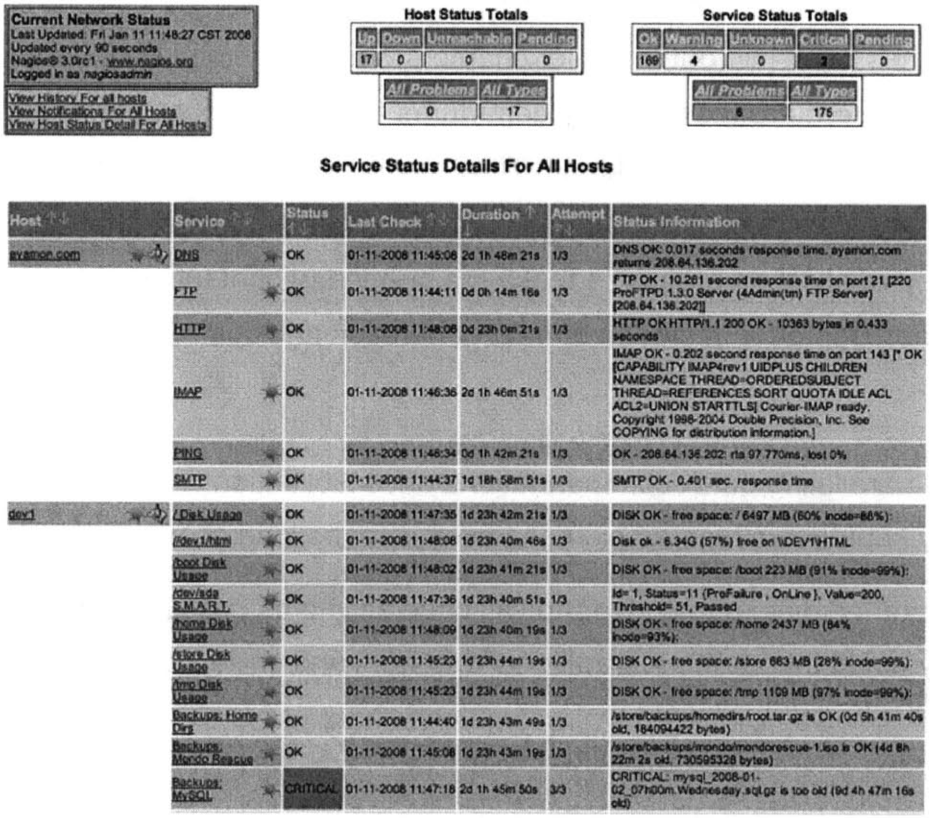
Production Websites (production-websites)

Host	Status	Services	Actions
ayamon.com	UP	6 OK	
nagios.com	UP	6 OK	
nagios.org	UP	6 OK	
nagioscommunity.org	UP	6 OK	

Switches (switches)

Host	Status	Services	Actions
linksys-srw224p	UP	1 OK	

Figure 26 Example Service Overview snapshot for Nagios



Current Network Status
Last Updated: Fri Jan 11 11:48:27 CST 2008
Updated every 90 seconds
Nagios® 3.0rc1 - www.nagios.org
Logged in as *nagiosadmin*

View History For all hosts
View Notifications For All Hosts
View Host Status Detail For All Hosts

Host Status Totals

Up	Down	Unreachable	Pending
17	0	0	0

All Problems	All Types
0	17

Service Status Totals

Ok	Warning	Unknown	Critical	Pending
169	4	0	2	0

All Problems	All Types
6	175

Service Status Details For All Hosts

Host	Service	Status	Last Check	Duration	Attempt	Status Information
ayamon.com	DNS	OK	01-11-2008 11:45:08	2d 1h 48m 21s	1/3	DNS OK: 0.017 seconds response time. ayamon.com returns 208.64.136.202
	FTP	OK	01-11-2008 11:44:11	0d 0h 14m 16s	1/3	FTP OK - 10.261 second response time on port 21 [220 ProFTPD 1.3.0 Server (4Admin(tm) FTP Server) [208.64.136.202]]
	HTTP	OK	01-11-2008 11:48:06	0d 23h 0m 21s	1/3	HTTP OK HTTP/1.1 200 OK - 10363 bytes in 0.433 seconds
	IMAP	OK	01-11-2008 11:46:36	2d 1h 46m 51s	1/3	IMAP OK - 0.202 second response time on port 143 [* OK [CAPABILITY IMAP4rev1 UIDPLUS CHILDREN NAMESPACE THREAD=ORDEREDSUBJECT THREAD=REFERENCES SORT QUOTA IDLE ACL ACL2=UNION STARTTLS] Courier-IMAP ready. Copyright 1998-2004 Double Precision, Inc. See COPYING for distribution information.]
	PING	OK	01-11-2008 11:46:34	0d 1h 42m 21s	1/3	OK - 208.64.136.202: rta 97.770ms, lost 0%
	SMTP	OK	01-11-2008 11:44:37	1d 18h 58m 51s	1/3	SMTP OK - 0.401 sec. response time
dev1	/ Disk Usage	OK	01-11-2008 11:47:35	1d 23h 42m 21s	1/3	DISK OK - free space: / 6497 MB (60% inode=88%):
	//dev1/html	OK	01-11-2008 11:48:08	1d 23h 40m 46s	1/3	Disk ok - 6.34G (57%) free on \\DEV1\HTML
	/boot Disk Usage	OK	01-11-2008 11:48:02	1d 23h 41m 21s	1/3	DISK OK - free space: /boot 223 MB (91% inode=99%):
	/dev/sda S.M.A.R.T.	OK	01-11-2008 11:47:36	1d 23h 40m 51s	1/3	Id= 1, Status=11 {PreFailure , OnLine }, Value=200, Threshold= 51, Passed
	/home Disk Usage	OK	01-11-2008 11:48:09	1d 23h 40m 19s	1/3	DISK OK - free space: /home 2437 MB (84% inode=93%):
	/store Disk Usage	OK	01-11-2008 11:45:23	1d 23h 44m 19s	1/3	DISK OK - free space: /store 663 MB (28% inode=99%):
	/tmp Disk Usage	OK	01-11-2008 11:45:23	1d 23h 44m 19s	1/3	DISK OK - free space: /tmp 1109 MB (97% inode=99%):
	Backups: Home Dirs	OK	01-11-2008 11:44:40	1d 23h 43m 49s	1/3	/store/backups/homedirs/root.tar.gz is OK (0d 5h 41m 40s old, 184094422 bytes)
	Backups: Mondo Rescue	OK	01-11-2008 11:45:08	1d 23h 43m 19s	1/3	/store/backups/mondo/mondorescue-1.iso is OK (4d 8h 22m 2s old, 730595328 bytes)
	Backups: MySQL	CRITICAL	01-11-2008 11:47:18	2d 1h 45m 50s	3/3	CRITICAL: mysql_2008-01-02_07h00m.Wednesday.sql.gz is too old (9d 4h 47m 16s old)

Figure 27 Example Service Details snapshot for Nagios

Chap. 5: Best Practices and Conclusions

Green HPC is a superset of generalized Datacenter design. We can therefore learn a considerable amount from not just the sustainable design construction community, with their consistently predictable electrical design improvements, but the "web farm" provisioning capabilities of the large Internet companies (Yahoo, Microsoft, Amazon, Facebook, Google, etc.) who have large data facilities purpose-built to deliver their web content. There are some novel approaches that represent promising paths in Datacenter design, but definitively, the best examples which can be applied to modern HPC Datacenter design are the Facebook Prineville Facility, which has an industry-wide (air-cooled) PUE low of 1.07 and came online in Spring 2011, Google's Datacenter "E" (which reported metrics in 2009 and 2010) and the NCSA Blue Waters Facility which has an HPC-specific Datacenter (water-cooled) low PUE of 1.16. It will be fully operational late this year or early 2012. Facebook has taken the unprecedented step of making all of their design documents Open Source materials available under a very loose Open Web Foundation license, which they call the Open Compute Project.

Highlights of the Open Compute Project

Much of which has been laid out in earlier chapters was adopted by Facebook in their new Prineville Facility. Here are their primary departures from traditional Datacenter design and their biggest "value-adds" to the dialogue. They managed to add value through subtraction in many cases on the design of their IT equipment. They removed items that aren't used in the fairly specialized application programs that a major Internet content provider uses. This removal also represented a decrease of heat generation and cost. Specifically, they used these strategies;

- They removed central UPS, opting instead for smaller, more efficient server-level UPS paralleling our recommendations of consolidation.
- They kept the voltage as high as possible for as long as possible. They eliminated a class of PDUs by utilizing 277VAC distribution at the IT equipment. This follows our contributed dictum to convert to DC once as late as possible in the power chain.
- They could then remove traditional ductwork and chillers because of the reduced energy footprint.

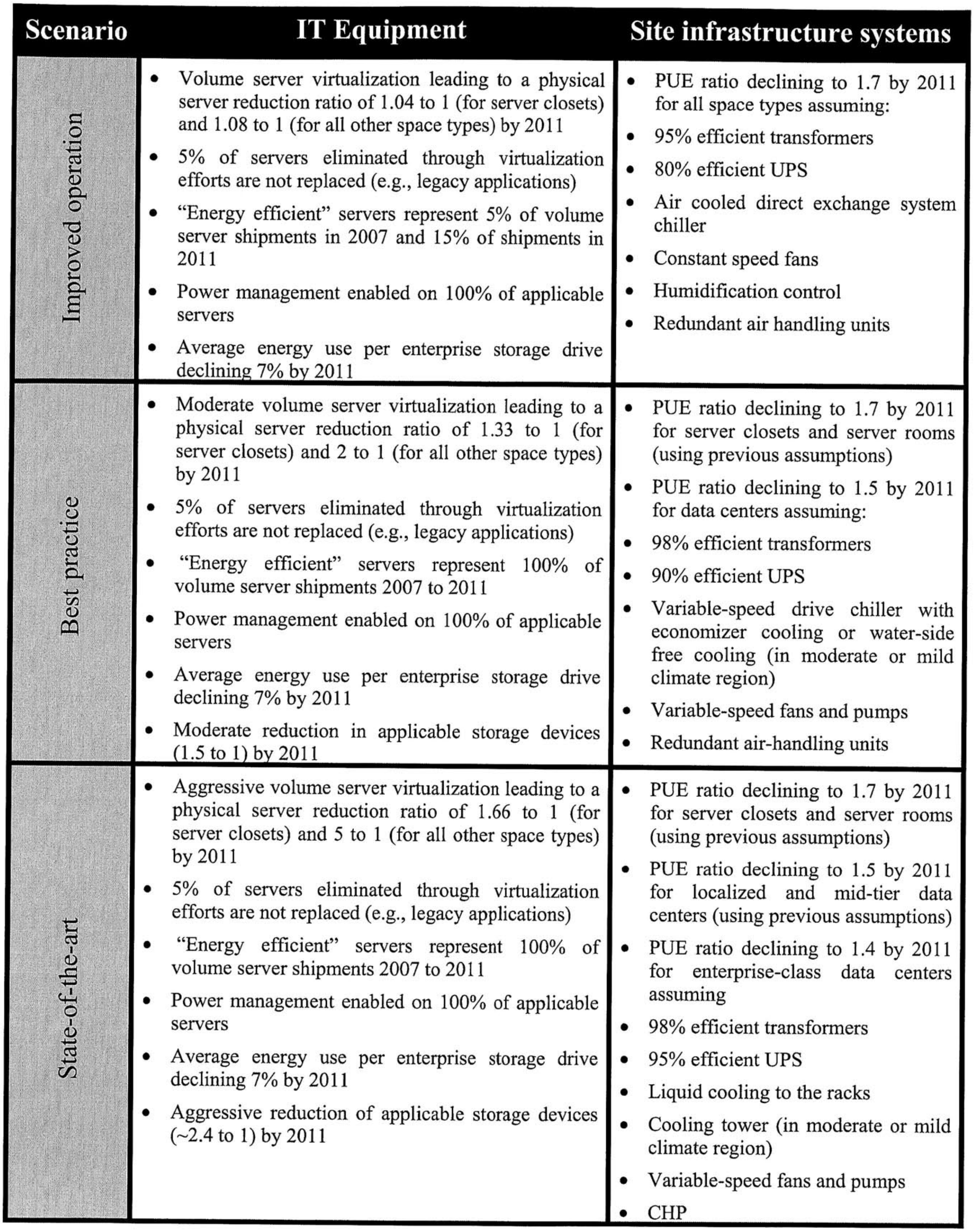
> *"The traditional data center loses 21% to 27% of its power due to inefficiencies built into the system. Losses enter the system during every stage of power transformation and conversion in a data center: When utility medium voltage is transformed to 480VAC, there is a 2% loss. Within the centralized UPS, there are two power conversions: AC to DC and DC back to AC, which results in a 6% to 12% loss. Power transformation at the PDU level from 480VAC to 208VAC results in a 3% power loss. Two-way server power supplies have two voltages: 208VAC to various DC voltages. This results in a 10% loss, assuming the power supply is an industry-average 90% efficient. By eliminating the centralized UPS and the PDUs, and by designing a 94.5% efficient server power supply, the Prineville data center has a total loss of 7.5% (including the 2% transformation loss). We relocated the UPS so it's closer to the server level and eliminates single points of failure upstream from the Open Compute servers in Prineville. Coupling this with the fact that we no longer need to synchronize between the centralized UPS and the PDUs, server availability increases from five 9s to six 9s (99.9999%)."*
> *(Park, Designing a Very Efficient Data Center, 2011)*

Google has one of the best industry PUEs (1.2) across the majority of their facilities. While there is considerable agreement in many of the underlying fundamentals of the Datacenter design recipe with the Open Compute approach, they do have a few construction guidelines that differ and we will highlight those here. For instance, we recommend utilizing the Google UPS approach (component-level as opposed to center-wide or rack-level) or none at all.

- Examine the entire cluster for a solution, not just the UPS
- Eliminate central UPS, distribute to each machine
- Eliminate AC-DC-AC double conversion
- Small battery, just enough to bridge until generators start up, or clean machine shutdown
- Incremental deployment of UPS capacity - no waste!
- Single-voltage PSU is the enabler, also made servers more power-efficient (not included in PUE computation)
- Real-world measured efficiency of >99.99%

Figure 28 from Insights into Google's PUE (Google, 2009)

These approaches are also espoused by Facebook, to which they attribute considerable power savings. Generally, motherboards use 2 voltages, 12v and 5v. If they only used 12v we could get rid of the step down transformers that power the 5v parts. These motherboards exist but are not in general availability. Google is working with manufacturers to make this happen in the future and to standardize on a single voltage motherboard.



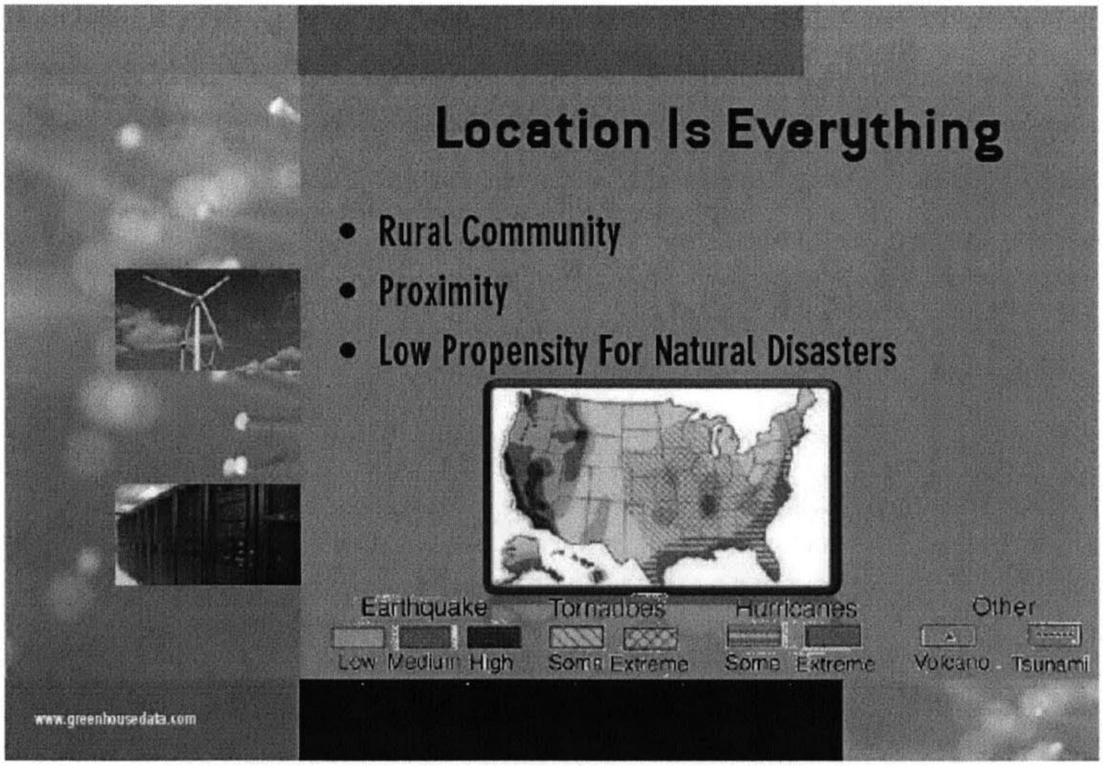
Scenario	IT Equipment	Site infrastructure systems
Improved operation	• Volume server virtualization leading to a physical server reduction ratio of 1.04 to 1 (for server closets) and 1.08 to 1 (for all other space types) by 2011 • 5% of servers eliminated through virtualization efforts are not replaced (e.g., legacy applications) • "Energy efficient" servers represent 5% of volume server shipments in 2007 and 15% of shipments in 2011 • Power management enabled on 100% of applicable servers • Average energy use per enterprise storage drive declining 7% by 2011	• PUE ratio declining to 1.7 by 2011 for all space types assuming: • 95% efficient transformers • 80% efficient UPS • Air cooled direct exchange system chiller • Constant speed fans • Humidification control • Redundant air handling units
Best practice	• Moderate volume server virtualization leading to a physical server reduction ratio of 1.33 to 1 (for server closets) and 2 to 1 (for all other space types) by 2011 • 5% of servers eliminated through virtualization efforts are not replaced (e.g., legacy applications) • "Energy efficient" servers represent 100% of volume server shipments 2007 to 2011 • Power management enabled on 100% of applicable servers • Average energy use per enterprise storage drive declining 7% by 2011 • Moderate reduction in applicable storage devices (1.5 to 1) by 2011	• PUE ratio declining to 1.7 by 2011 for server closets and server rooms (using previous assumptions) • PUE ratio declining to 1.5 by 2011 for data centers assuming: • 98% efficient transformers • 90% efficient UPS • Variable-speed drive chiller with economizer cooling or water-side free cooling (in moderate or mild climate region) • Variable-speed fans and pumps • Redundant air-handling units
State-of-the-art	• Aggressive volume server virtualization leading to a physical server reduction ratio of 1.66 to 1 (for server closets) and 5 to 1 (for all other space types) by 2011 • 5% of servers eliminated through virtualization efforts are not replaced (e.g., legacy applications) • "Energy efficient" servers represent 100% of volume server shipments 2007 to 2011 • Power management enabled on 100% of applicable servers • Average energy use per enterprise storage drive declining 7% by 2011 • Aggressive reduction of applicable storage devices (~2.4 to 1) by 2011	• PUE ratio declining to 1.7 by 2011 for server closets and server rooms (using previous assumptions) • PUE ratio declining to 1.5 by 2011 for localized and mid-tier data centers (using previous assumptions) • PUE ratio declining to 1.4 by 2011 for enterprise-class data centers assuming • 98% efficient transformers • 95% efficient UPS • Liquid cooling to the racks • Cooling tower (in moderate or mild climate region) • Variable-speed fans and pumps • CHP

Figure 29: EPA's Summary of Alternative Efficiency Scenario Assumptions from 2007

Google and Facebook have made improvements better than those predicted by the EPA in the 2007 report (see Table above). The EPA appears to have predicted conservatively since our exemplar centers have outstripped even their most aggressive predictions.

Locating Datacenters

There are not many locations in the US that are on the "smart grid" yet so there will be no way for you to benefit financially from any efficiencies you can bring in upstream. But you can build where the conditions are optimal for large Datacenters. There are certain Datacenters, notably Green House Data, who have located to the places where there is cheap electricity. Wyoming is cold and windy and they give substantial incentives to get businesses to relocate there. Not coincidentally it is relatively immune to natural disasters. These location decisions are key if you plan on doing on-site power generation since you may be able to use renewable energy resources if they are available in the location you choose as is the case with Green House Data, who uses a co-located wind farm for part of their energy needs. From an efficiency perspective it is not important to put the Datacenter physically near the customer. Put it where it is cheapest and coolest and the customer can access it remotely. This has been a convention in HPC for many years.

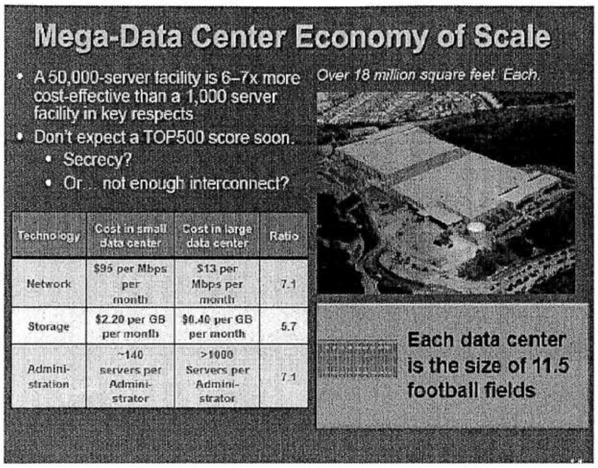


Location Is Everything

- Rural Community
- Proximity
- Low Propensity For Natural Disasters

Figure 30: Geographical assessment of potential Datacenter locales
(Mills, 2010)

Learning from the big farms; Google's low PUE.



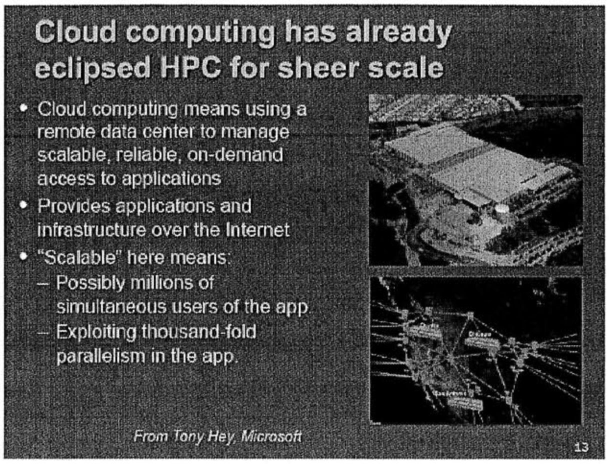
Mega-Data Center Economy of Scale

- A 50,000-server facility is 6–7x more cost-effective than a 1,000 server facility in key respects
- Don't expect a TOP500 score soon.
  - Secrecy?
  - Or... not enough interconnect?

Technology	Cost in small data center	Cost in large data center	Ratio
Network	$95 per Mbps per month	$13 per Mbps per month	7.1
Storage	$2.20 per GB per month	$0.40 per GB per month	5.7
Admini-stration	~140 servers per Admini-strator	>1000 Servers per Admini-strator	7.1

Over 18 million square feet. Each.

Each data center is the size of 11.5 football fields



Cloud computing has already eclipsed HPC for sheer scale

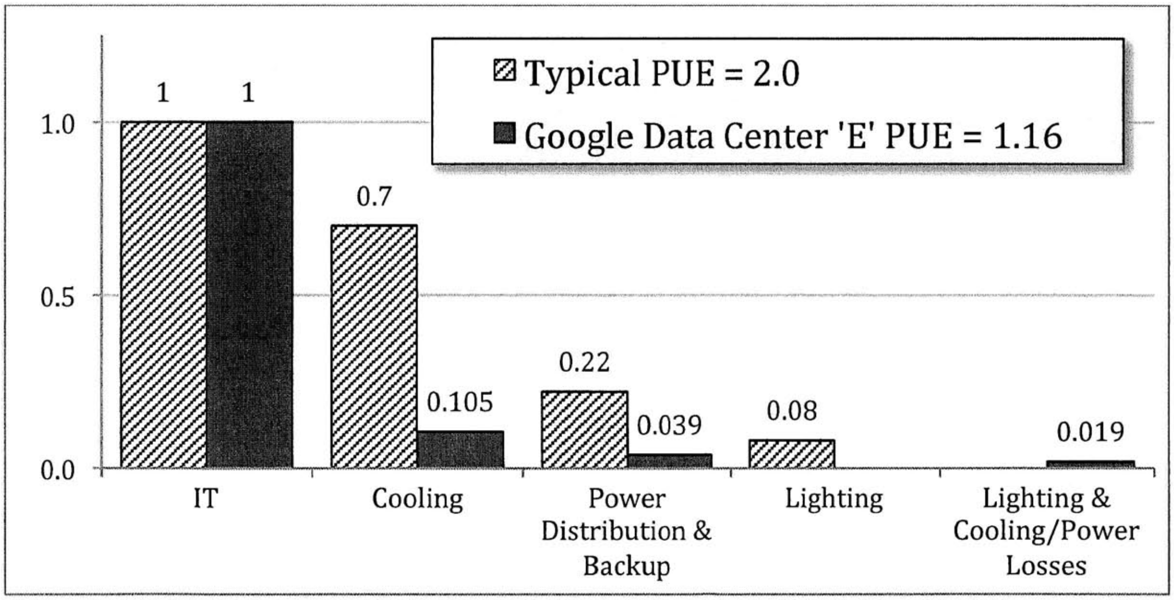
- Cloud computing means using a remote data center to manage scalable, reliable, on-demand access to applications
- Provides applications and infrastructure over the Internet
- "Scalable" here means:
  - Possibly millions of simultaneous users of the app.
  - Exploiting thousand-fold parallelism in the app.

From Tony Hey, Microsoft

13

Figure 31 and Figure 32 The Scale of Data Centers (Gustafson, 2009)

Google has substantially lowered its PUE by doing away with a majority of the power draws that cannot be tracked to useful work. They made efficiency improvements throughout the Datacenter pyramid to cascade power savings up the chain resulting in a considerable reduction in cooling requirements. Below are 2 bar chart representations of the Google PUE comparison. A bar represent a percentage of the power usage. For instance, cooling in a traditional center is 70% of the overall power use while it is close to 10% in the Google model.



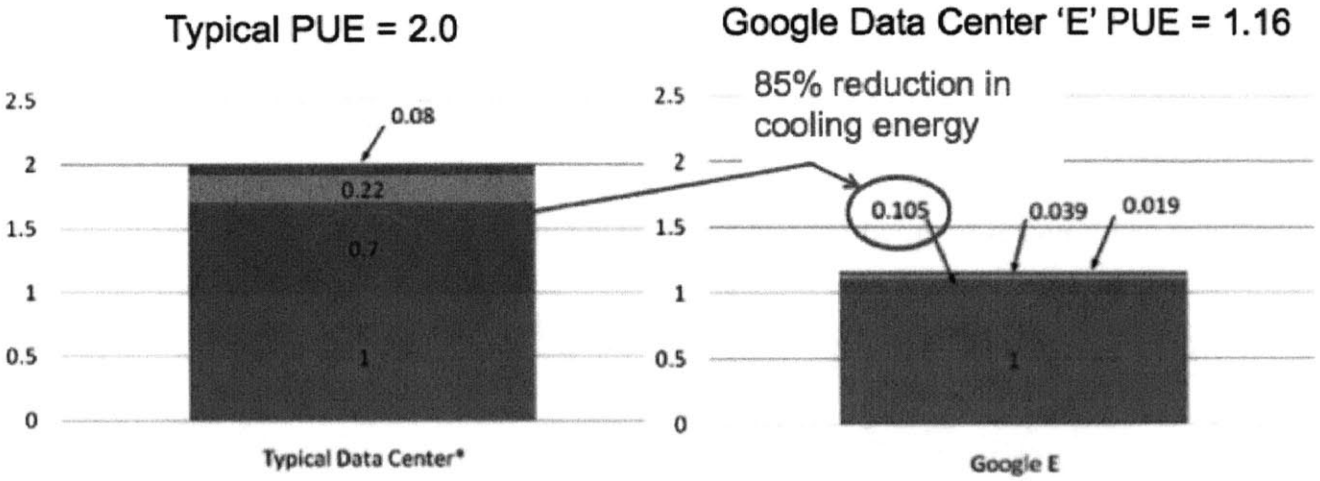


Figure 33 Typical PUE, Google Data Center PUE (Google, 2009)

Facebook runs their servers in a hotter environment than traditional centers because their servers can tolerate those higher temperatures. And the air handling within those servers has been made more efficient through their innovative "triplet" design.

Below is a typical Psychometric chart used in determining the efficiencies of the Datacenter temperature targets in your intakes, exhausts, and production room. They have demonstrated that you can run your servers at 70 – 80 degrees F in the cold aisle. These temperatures are tolerated by modern server equipment and can significantly reduce your need for cold air or a large chiller system.

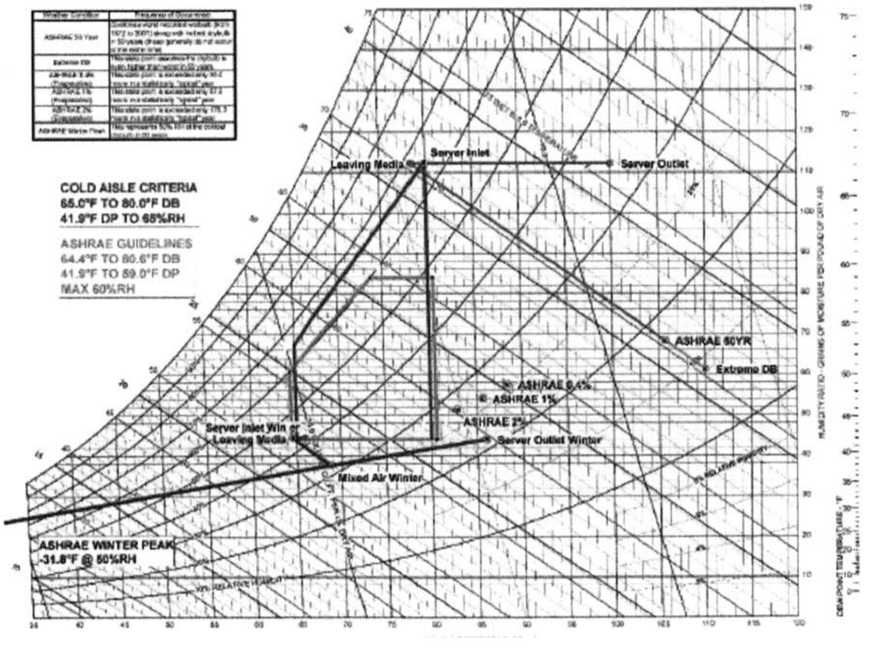
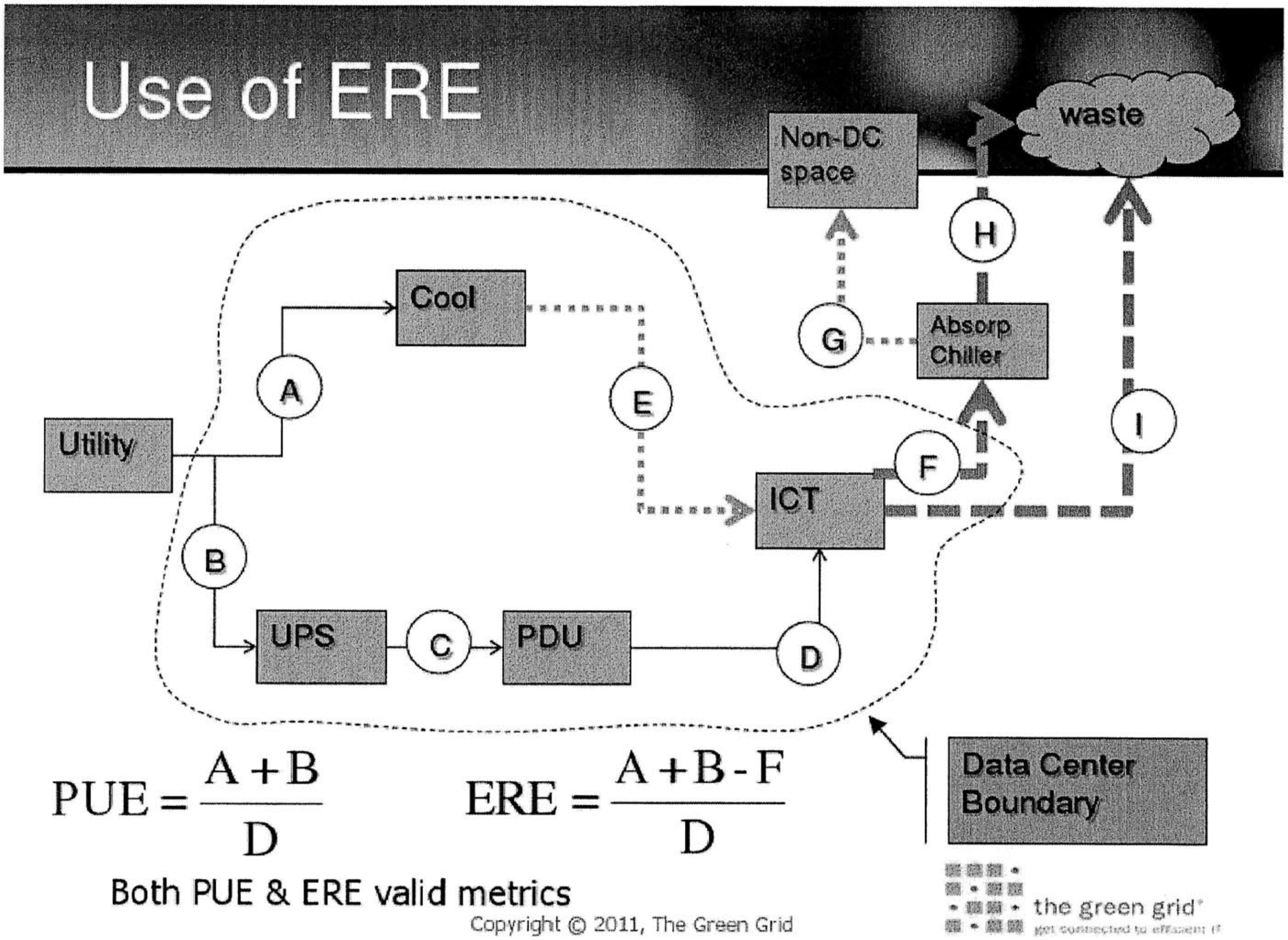


Figure 34 Prineville Psychometric chart. (Park, Data Center v1.0, 2011)

The new metrics: capturing efficiency all through the chain

We need to use better metrics to describe and characterize our energy-efficiencies. The Green Grid, who originally formulated the PUE, suggests a new metric, which they call Energy Reuse Effectiveness (ERE). This calculation allows you to capture some of the value associated with reusing energy to condition occupied workspaces.



Use of ERE

$$PUE = \frac{A+B}{D} \qquad ERE = \frac{A+B-F}{D}$$

Both PUE & ERE valid metrics

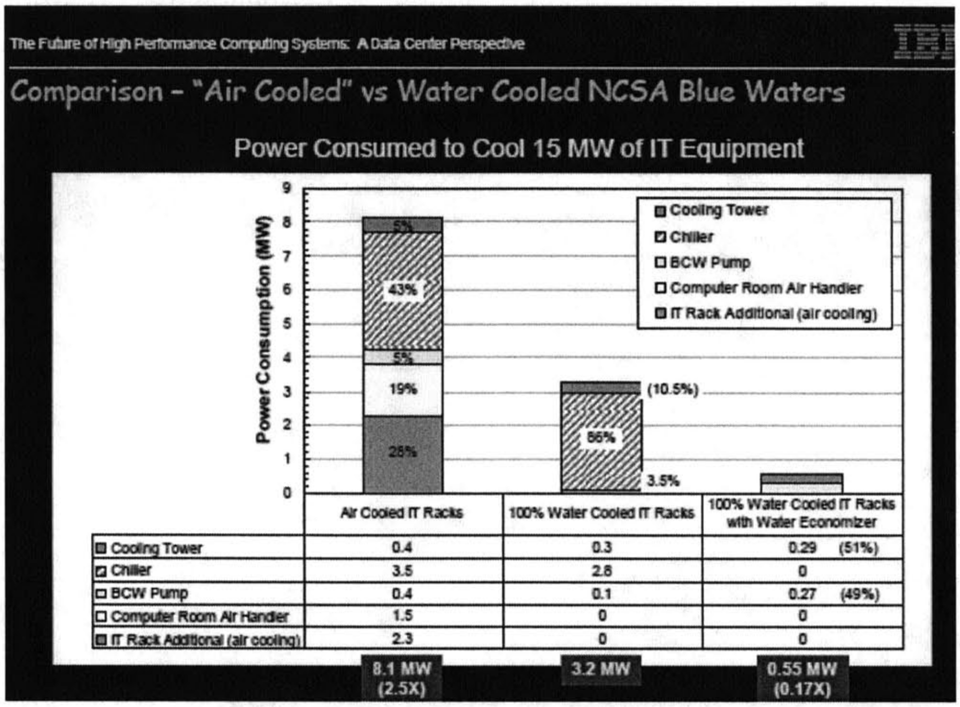
Copyright © 2011, The Green Grid

Figure 35 Energy Reuse Effectiveness (Azevedo, Cooley, Patterson, & Blackburn, 2011)

ERE is a better, more comprehensive tool than PUE (in draft standard)

ERE = (Total - Reused) / ICT = (Cooling + Power + Lighting + IT – Reused) / ICT

Water cooling Best Practice (NCSA Blue Waters)



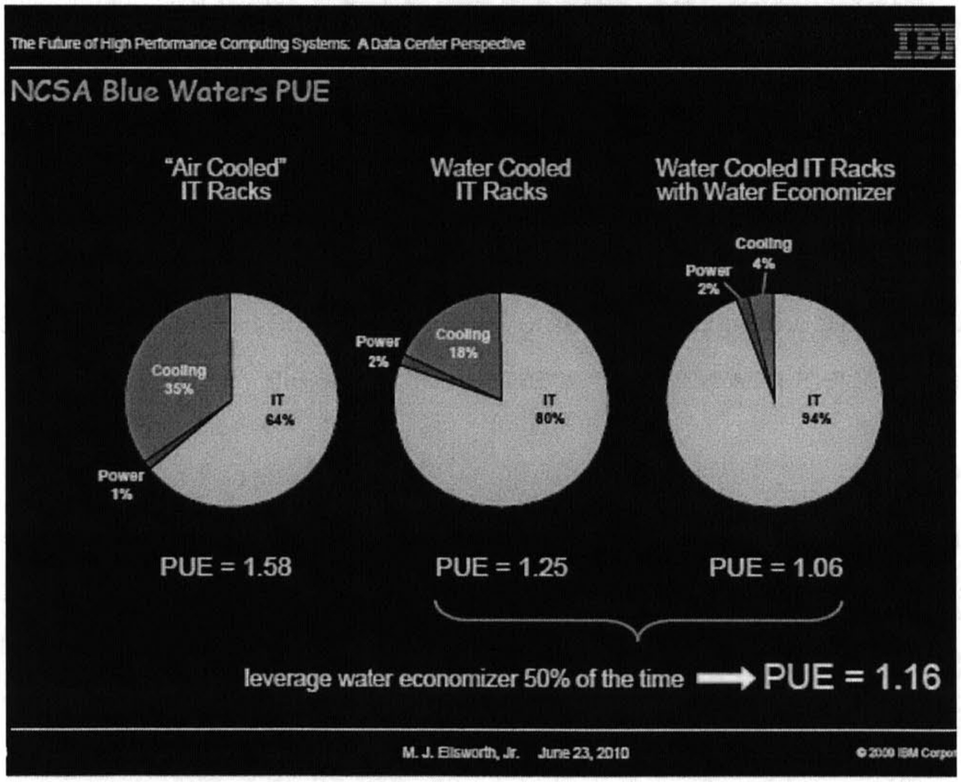
The Future of High Performance Computing Systems: A Data Center Perspective

Comparison – "Air Cooled" vs Water Cooled NCSA Blue Waters

Power Consumed to Cool 15 MW of IT Equipment

	Air Cooled IT Racks	100% Water Cooled IT Racks	100% Water Cooled IT Racks with Water Economizer
■ Cooling Tower	0.4	0.3	0.29 (51%)
▨ Chiller	3.5	2.8	0
□ BCW Pump	0.4	0.1	0.27 (49%)
□ Computer Room Air Handler	1.5	0	0
■ IT Rack Additional (air cooling)	2.3	0	0
	8.1 MW (2.5X)	3.2 MW	0.55 MW (0.17X)

Figure 36 Data Direct Networks



The Future of High Performance Computing Systems: A Data Center Perspective

NCSA Blue Waters PUE

PUE = 1.58

PUE = 1.25

PUE = 1.06

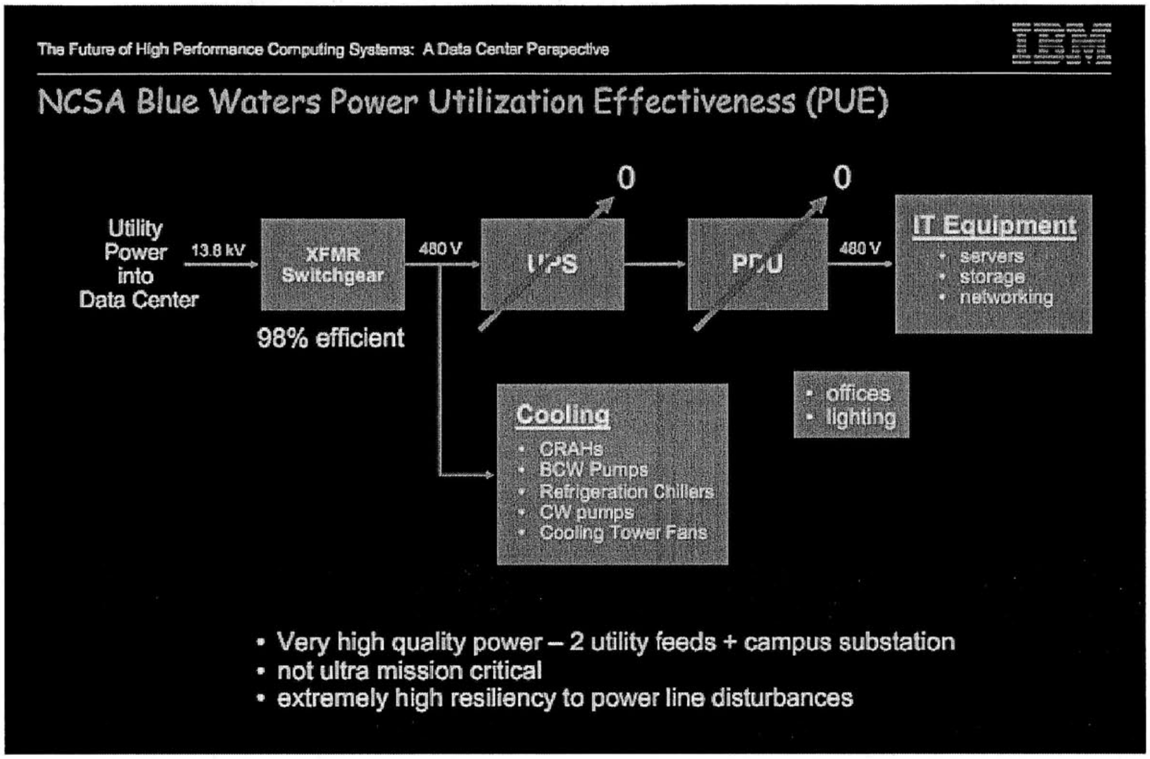
leverage water economizer 50% of the time → PUE = 1.16

M. J. Ellsworth, Jr. June 23, 2010

© 2009 IBM Corpo

Figure 37: Blue Waters PUE (Ellsworth, 2010)

The DOE funded Blue Waters facility, which is located at the National Center for Supercomputing Applications at the University of Illinois, will be a 10 PFLOPs facility when it is benchmarked early next year. The significant discussion point they bring to the dialogue is their substantial PUE reduction achieved through their construction design around water cooling. It is notable that they have a PUE of 1.06 when they leverage server-side economizing.



The Future of High Performance Computing Systems: A Data Center Perspective

NCSA Blue Waters Power Utilization Effectiveness (PUE)

Utility Power into Data Center

13.8 kV

XFMR Switchgear

98% efficient

480 V

UPS

0

PDU

0

480 V

IT Equipment
- servers
- storage
- networking

Cooling
- CRAHs
- BCW Pumps
- Refrigeration Chillers
- CW pumps
- Cooling Tower Fans

- offices
- lighting

- Very high quality power – 2 utility feeds + campus substation
- not ultra mission critical
- extremely high resiliency to power line disturbances

Figure 38 Blue Waters PUE

Much in the same way as Google and Facebook have done, they have removed centralized UPS and PDUs to keep voltage at the same level down to the IT server equipment with its resultant energy-efficient power savings cascade up the chain.

Future trends

Selection of alternate processor technologies

The path to exascale supercomputing does not go through the current Top500 List. We have seen that the current trends at the top of that rating system are unsustainable. The future of Datacenter design can be found in the Green500 list but the exact path must be teased out of the data. (Green500.org, 2010)

Green500 Rank	Top500 Rank	Site	Computer
1	115	IBM Thomas J. Watson Research Center	IBM NNSA/SC Blue Gene/Q Prototype
2	4	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, NVIDIA GPU, Linux/Windows
3	404	NCSA	Hybrid Cluster Core i3 2.93Ghz Dual Core, NVIDIA C2050, Infiniband
4	170	RIKEN Advanced Institute for Computational Science	Fujitsu K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect
5	207	Forschungszentrum Juelich (FZJ)	IBM QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus
5	208	Universitaet Regensburg	IBM QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus
5	209	Universitaet Wuppertal	IBM QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus
8	22	Universitaet Frankfurt	Supermicro Cluster, QC Opteron 2.1 GHz, ATI Radeon GPU, Infiniband
9	117	Georgia Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5660 2.8Ghz, NVIDIA Fermi, Infiniband QDR
10	102	National Institute for Environmental Studies	GOSAT Research Computation Facility, NVIDIA GPU
11	1	National Supercomputing Center in Tianjin	NUDT YH Cluster, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C

Top 11 entries in the Green500 list.

9 of the 11 most efficient Datacenters in the Green500 are accelerated with either GPUs or IBM PowerXCells. The accelerator approach helps to solve the Power Wall problem. Also well-represented high up in the Green500 list is the IBM Blue Gene design, which has successfully worked around the Memory Wall. The trend indicates that getting away from general-purpose CPUs and espousing a “reverse-acceleration” model, which was first demonstrated on the LANL Roadrunner supercomputer, will get us both performance and energy-efficiency gains. (Pakin, Lang, & Kerbyso, 2009)

Onsite Power Generation

In addition to the wind power supported Datacenters discussed earlier, there are a growing number of centers that are incorporating solar power. This will be an increasing trend as the price of solar panels continues to drop. This too, is determined by an analysis of the CAPEX cost of this piece of the infrastructure relative to its displacement of VAREX.

> *"DataScan Technologies which ... opened a facility in Alpharetta, Georgia ... says its solar array ... will produce an estimated 285,500 kWh hours of electricity annually, reducing carbon emissions by 205 metric tons. The company says the solar power will reduce its data center's energy consumption by 10 percent in 2011." (Miller, 2011)*

Additional approaches and concerns

There is considerable experimentation with placing IT equipment in shipping containers to take additional control over the cooling and cabling associated with tightly packed servers. This approach has taken lessons from the contemporary literature to heart and has implemented many of the best practices we have defined in an effort to achieve a high degree of efficiency. Many businesses are looking at this solution, making the replaceable element of the Datacenter a whole row of racks, rather than a single rack or subcomponent within a server.

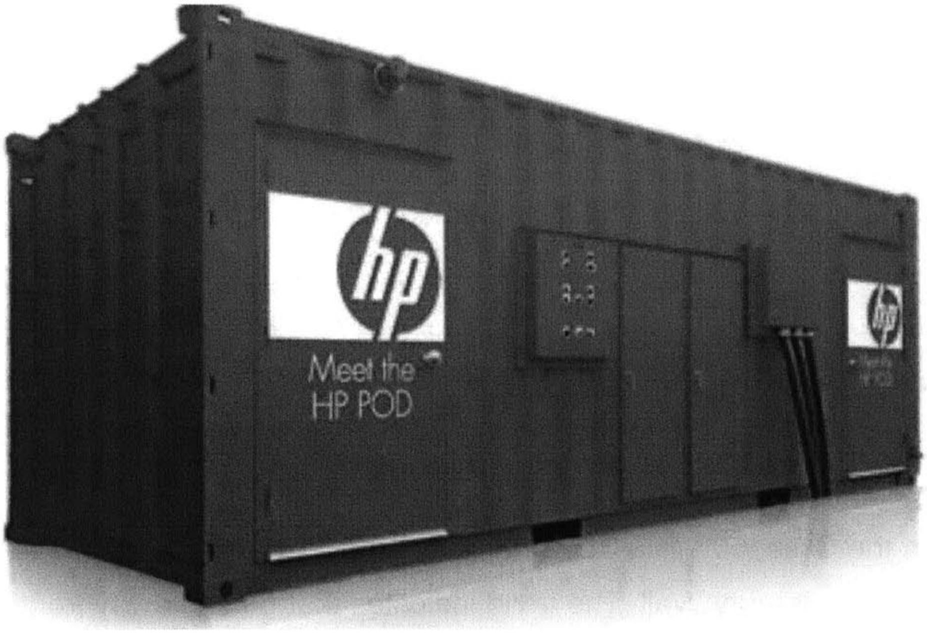


Figure 39: A typical containerized Datacenter (Ramanunni, 2011)

Conclusion: The systems approach is the path to Green HPC

We have seen, through an analysis of the approach of a number of enlightened Datacenter designs as well as through the proper implementation of our contributed power usage reduction recommendations, where one can gain energy efficiencies at every step of the HPC chain. If you synchronize all of these improvements you will get a sizeable increase in efficiency and a drastic power reduction. After you have performed a bottom-up analysis, you can hand off a much reduced power budget to the facilities planners to make their cooling calculations.

Since the release of the EPA report in 2007, Datacenters have improved their efficiencies across the enterprise to get to a very respectable PUE. The theoretical low PUE is 1, and many Datacenters have implemented designs and practices that substantially get us most of the way there.

Year	Site	PUE	# of Centers in Sample	Source
2006	Industry Average	2.00	22	(Fanara, et al., 2007)
2008	DOE Average	1.47	41	(Energetics Incorporated, 2010)
2009	DOE Average	1.44	49	(Energetics Incorporated, 2010)
2010	Google Datacenter E	1.20	1	(Google, 2009)
2011	Facebook Prineville	1.07	1	(Park, Data Center v1.0, 2011)
2012	NCSA Blue Waters	1.16*	1	(Ellsworth, 2010)

* anticipated

We can cross-reference our exemplar Datacenters across the 10 Rules of Energy Efficient Design to show what areas they have innovated in and how well it lines up with our contributed recommendations (on a range of Low, Mid, and High).

Site	The operation not performed is the most energy-efficient.	Implement HPC equipment maintainer / user behavior modifications.	Consolidation and distribution to leverage hybrid architectures.	Give the big problems their due emphasis but also solve the lots of little problems.	Find a way to effectively utilize idle cycles for computation.	Compile code locally to maximize resource usage.	Use the most numerically efficient approach.	Convert to DC once and stay there.	Uncouple PE to RAM consolidation to ameliorate the Memory Wall.	Overlap and integrate computation and communications.
Industry Average	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
DOE Average	Low	Mid	Low	Low	Mid	Low	Low	Low	Low	Low
DOE Average	Mid	High	Mid	Low	Mid	Mid	Low	Low	Low	Low
Google Datacenter E	High	NA	High	High	High	NA	Mid	Mid	Mid	Mid
Facebook Prineville	High	NA	High	High	High	NA	Mid	High	Mid	Mid
NCSA Blue Waters	High	High	High	High	High	High	Mid	High	Mid	High

Additional improvements in efficiencies can be gained through proper application of the guidelines prescribed in this document. While the designers of the Datacenters exemplifying the best practices in this field have analyzed the whole chain of power use in their respective centers, there are still opportunities to improve as new equipment is released and new methods are implemented in software and user behavior management.

We have seen how the large web content and Internet services companies have learned from best practice recommendations in Datacenter construction and are incorporating the 10 Rules of Energy Efficient Design into their ongoing development efforts.

> *"Large-scale data-intensive applications, such as high performance key-value storage systems, are growing in both size and importance; they now are critical parts of major Internet services such as Amazon, LinkedIn, and Facebook." (Andersen, 2009)*

In the future these tenets will be even more important to keep in focus as new directions become available, especially in HPC. These opportunities may come in the form of new power options. Perhaps there will be an even more right-sized power distributions standard than the 277VAC standard that Facebook has built upon. If that becomes available, Datacenter designers will need to run the value-chain analysis again to determine their resultant VAREX costs and see if that equipment modification makes sense for them. It is likely we will see more use of DC in future designs because of the demonstrated advantages in getting your current to the servers in the form that is most efficient for them. Additional technologies that are just on the horizon for Datacenters are the further developments predicted for parallel programming languages. As we have seen, incremental improvements in how your algorithms are optimized for performance, and are then compiled for the same, can reveal substantial local energy savings and then will cascade up the chain to give you a substantially lower overall power budget.

The system engineering approach used to analyze and optimize Datacenter design is a considerable asset in determining areas of energy-efficiency. Further, it has proven to be a useful tool to expose systemic effects up the power chain and to guide us to a solution that represents the optimum Green HPC energy savings.

Glossary of HPC Definitions

High Performance Computing (HPC) is characterized by the following "laws," observations, and terminology. To a great extent these have been the coin of the realm since the beginning of HPC and will certainly be the standards to which future "Peak Performance" and "Green" HPC will be held to. A full understanding of the terms used below will lead one to an appreciation of the desires and limits of the HPC design world.

10GE and 10GigE

10 Gigabit per second Ethernet (IEEE Std. 802.3ae-2002).

ACPI

Advanced Configuration and Power Interface is a specification that provides a standard for device configuration and power management by the operating system.

AHCI

Advanced Host Controller Interface is a standard defined by the Intel Corporation that specifies the operation of hard drive bus adapters.

Amdahl's Law

Amdahl's law defines the maximum speedup available from an algorithm on a particular system. It holds because parallel algorithms almost always include work that can only take place sequentially. From this sequential fraction, Amdahl's law provides a maximum possible speedup. Amdahl's Law has been obviated to a great extent by the way modern HPC systems are constructed. No longer are we limited to a single path for any subsystem (processor, memory, network).

API

The Application Programming Interface is a set of rules that programs can follow to communicate with each other through a shared reference library.

Bandwidth

A measure of data communications usually represented in bits or bytes per second. This measurement is used in describing the capabilities of virtually every subsystem within the supercomputer model; communication speeds between processor and memory systems, speeds between processors and storage devices, and network communication between compute nodes.

Batch or Job Scheduler

A software application that is in charge of unattended background executions.

BIOS

The Basic Input/Output System is a low-level set of instructions, usually in a server's firmware, that fix certain configuration settings prior to the Operating System booting.

BLAS

The Basic Linear Algebra Subprograms (BLAS) is a standard for publishing libraries to perform basic operations such as vector and matrix multiplication.

CFM — Cubic feet per minute.

CISC — Complex Instruction Set Computing

See RISC

Cluster Computing

In computers, clustering is the use of multiple computers, typically PCs or UNIX (Linux) workstations, multiple storage devices, and redundant interconnections, to form what appears to users as a single highly available system. Cluster computing can be used for load balancing as well as for high availability. Advocates of clustering suggest that the approach can help an enterprise achieve 99.999% availability in some cases. One of the main ideas of cluster computing is that, to the outside world, the cluster appears to be a single system (a Supercomputer).

CMOS

Complementary Metal–Oxide–Semiconductor is a technology for constructing integrated circuits. This term is often used as a replacement for BIOS.

Convective cooling

Heat transfer by natural, upward flow of hot air from the device being cooled.

Cooling Economizer

The use of outside air to keep the Datacenter cool.

CPE – Customer Premises Equipment

"Customer Premises Equipment" is a catch-all term for gear owned by the electric utility but is physically located at the Datacenter. This may come in the form of a high-voltage subsystem located adjacent to the Datacenter, a "Meet-me" room inside the building where the utility company meters its use, or other point of connection and power delivery from the electric company. This marks our starting point for the energy efficiency analysis.

CPU

The Central Processing Unit is the core of a computer and carries out the instructions of a computer program.

CRAC and CRAHs — Computer Room Air Conditioner and Computer Room Air Handler

These are a necessary part of today's Datacenter. The Emerson / Liebert company has the majority of this market. Their products are often measured in tons (which is a rating indicating how much airflow they can handle) or kilowatts. Generally speaking, you want to oversize the air conditioner in your Datacenter when you build it because since this is a piece of the infrastructure you will not have the opportunity to put a larger unit in down the road. Most Datacenters don't oversize it by much so in the Datacenter's operational life you need to track the power-using additions to the Datacenter to ensure that they aren't outstripping your electrical supply or your heat removal capabilities.

DARPA

Defense Advanced Research Projects Administration

Datacenter

A Datacenter is a specialized facility designed specifically for housing servers. These servers can be used for Internet services like website hosting or for tightly bound scientific computation like the programs found in HPC.

DCiE

Data Center Infrastructure Efficiency is a metric that is the inverse of PUE.

Disk

The hard drive component of a PC or storage subsystem.

Distributed computing

Systems like SETI@Home claim to have hit a PetaFLOPS a number of years ago, but the latency associated with it is considerable, so it is only a good approach for certain classes of problems that can tolerate high latencies like the Mersenne Prime Number search. Still, the various @Home and Distributed computing projects are notable for their historical performance. Protein Folding problems are one of the popular research areas in Biological Engineering, attempting to discover the causes of diseases like Muscular Dystrophy and ALS. Some of these areas are being investigated by Folding@Home. Other Protein Folding problems are being investigated by tightly-coupled low latency solutions like the Anton Supercomputer from DRShaw Research.

DOD

Department of Defense

DOE

Department of Energy

Double Precision

The IEEE 754-2008 Double Precision standard is a technical standard that describes the floating point math standard to which computational infrastructure manufacturers must comply in order for code to be portable between systems. Compliance with this standard is required to get your system rated on the Top500 list. The adoption of this standard has had considerable controversy around it.

DVFS

Dynamic Voltage and Frequency Scaling, which describes the capacity to slow down processors to use less electricity during idle periods.

ECC - Error-Correcting Code

ECC is used as a hardware technique to better guarantee the accuracy of bytes of memory. ECC RAM is a type of memory that includes special circuitry for testing the accuracy of data as it passes in and out of memory. This functionality comes with some overhead cost which is not always necessary.

EEE

A data switch subscribing to the IEEE P802.3az standard.

Embarrassingly parallel

This term describes an algorithm or application that scales linearly with the amount of resources applied to it. This describes the majority of HPC problems in the field today and is well-suited to the cluster approach used by HPC centers today.

Energy Star

An international standard for energy efficient consumer products.

EPA

Environmental Protection Agency

Ethernet

A data communications standard common in HPC.

Exascale

A class of Supercomputers that can generate ExaFLOPS. This is a target set by the HPC community in order to perform certain classes of modeling and simulation that we cannot currently do at our existing centers.

FFT

A Fast Fourier Transform is an efficient algorithm used to compute the discrete Fourier transform (DFT) and its inverse. It is often used to benchmark HPC systems.

FLOPS (alternately FLOP/s), OPS or IPS

Floating Point Operations per Second. This (or sometimes "Instructions Per Second" for non-math intensive operations) is the number we use to compare supercomputing systems performance. It is the metric used on the Top500 list to show raw performance. On the Green500 list it is in the numerator to show energy efficient supercomputer ratings (and power is in the denominator).

Gigabit Ethernet or GigE

An Ethernet standard that runs at 1 Gigabit per second as defined by the IEEE 802.3-2008 standard.

GPU

A Graphics Processing Unit is a co-processor used for offloading computational work on many modern HPC systems.

Green computing or green IT

This refers to environmentally sustainable computing or IT. In the article Harnessing Green IT: Principles and Practices, San Murugesan defines the field of green computing as "the study and practice of designing, manufacturing, using, and disposing of computers, servers, and associated subsystems ... efficiently and effectively with minimal or no impact on the environment." (Murugesan, 2008)

Green HPC

The extension of Green IT techniques across an enterprise Datacenter. Many individual improvements can be enabled on the intrinsic elements of the Datacenter that can then drive improvements on the building-level energy saving efforts.

Green500.org

The Green500 list repackages the Top500 list in the numerator divided by the power usage of those Datacenters in the denominator. This gives us an energy efficiency rating in a reordered top 500 list.

Grid computing

Grid computing is the combination of computer resources from multiple administrative domains for a common goal.

HPC — High-Performance Computing

The use of parallel programming for running advanced application programs sufficiently reliably and quickly. Most HPC centers these days operate at a level above a teraflops. The highest performing HPC centers operate at the petaflops levels. For the purposes of this paper "HPC" and "supercomputer" terms are used interchangeably.

HVAC

Heating, Venting, and Air Conditioning

Hypervisor

A hardware virtualization technique that allows multiple operating systems, termed guests, to run concurrently on a host computer.

IEEE

The Institute of Electrical and Electronics Engineers is the largest professional organization in the HPC space and has had considerable influence on defining the standards to which HPC designers comply.

Infiniband

A switched fabric communications link used in high-performance computing and enterprise data centers. Its features include high throughput, low latency, quality of service and failover, and it is designed to be scalable.

I/O

Input / Output. A catch-all term for data communications.

Latency

The measurement of time delay in communications. Within the packet switched network subsystems in the Datacenter there are quite a few sources that can introduce latency into our computation chain throughout the data processing paradigm.

LEED — Leadership in Energy and Environmental Design

LEED is a green building certification system managed by the US Green Building Council. A particular level of certification is awarded to a building which complies with a number of energy-efficient design criteria. A LEED standard for Datacenters has been proposed by a group based out of Lawrence Berkeley National Laboratory. It is an environmental performance criteria (EPC) guide for new construction ((LBNL), 2009). Note that there is currently not a comprehensive guide for retrofitting old Datacenters.

LINPACK

A software library used for performing numerical benchmarks on computers. This library, or a tool called high-performance LINPACK, is used to rate and then compare different supercomputing implementations. It is the single most important benchmark in HPC and is required for consideration to be placed on the Top500 list. There is some disagreement surrounding the selection of LINPACK as the benchmark for HPC. Many noted scholars, including Intel's Gustafson, believe we should use the Fast Fourier Transform Benchmark. Both of these are part of the full HPC Challenge benchmark suite.

Megawatt (MW)

1,000,000 Watts. Most Datacenters draw MWs of power.

Memory Wall

See Von Neumann Bottleneck

Metcalfe's law

This states that the value of a telecommunication network is proportional to the square of the number of connected users. This generally predicts that the utility of the system goes up as you add users or processing elements.

MIMD — Multiple Instruction Multiple Data processing

Processing that allows us to perform multiple arithmetic operations on a piece of data while it is in memory. Moving data in and out of memory is where you take quite a few of our energy efficiency hits. It has been the focus of considerable effort to reduce memory reads and writes and therefore power use.

MOG

MIMD On a GPU

Moore's Law

More an observation than a law, this describes the phenomenon that the number of transistors that can be placed inexpensively on integrated circuits doubles approximately every 18 months. CPUs would shrink in size in the 1990s but nowadays the die size is holding firm while we are doubling the number of cores in a CPU.

Motherboard

A motherboard is the central printed circuit board (PCB) in many modern computers and holds many of the crucial components of the system, while providing connectors for other peripherals.

MPP — Massively Parallel Processing

MPP is the coordinated processing of a program by multiple processors that work on different parts of the program, with each processor using its own operating system and memory. This term describes much of supercomputing today. Because of the way computer systems are designed, dividing the work load up over many computational cores paired with local memory has proven to be the way to substantially accelerate and improve computational performance.

Network

A computer network is a collection of computers and devices interconnected by communications channels that facilitate communications and allows sharing of resources and information among interconnected devices.

PDU

Power Distribution Unit is a broad term that describes the equipment that receives electricity in and breaks it into the constituent components and voltages that we need for the Datacenter equipment. In our usage, it generally takes it from the "mains" at a high voltage and breaks it up into multiple 110 or 120 volt AC service for the servers.

PEs

Processing elements, usually a CPU core or virtualized processor.

Petascale

A class of Supercomputers that can generate PetaFLOPS (10^{15} FLOPS). This is where the state-of-the-art HPC Datacenters are at today. Without exception, the Petascale systems that we know of in the US have been funded through a series of DOE programs, notably ASCI and ASCR. We have substantially arrived at this state of the art using the techniques that leverage Moore's Law.

Power Wall

This is the phenomenon that tells us we will hit a power utilization ceiling before we hit the physical limits of Moore's Law. This is the Number 1 enemy of Green HPC. It has been predicted that we will no longer be able to construct CPUs as we are doing it now by the end of the decade and will need to find an alternate approach in order to get to exascale computing.

Precision

Precision, in computer programming parlance, describes the data representation of numbers. It is usually either single or double. A single precision number takes up 32 bits in memory while a double precision number takes up 64 bits. Sometimes additional bits are used for error checking. This will turn out to be significant when we discuss some of the data acceleration options. As we will see, it is more expensive to process double precision registers in the CPUs and GPUs than it is to make registers capable of storing single and half precision numbers and doing less double precision math in a process known as iterative refinement.

PS or PSU — Power Supply Unit

The Power Supply on the server. This converts 120 or 240v AC to a number of different DC voltages.

PUE —

Power Usage Effectiveness is a metric used to determine the energy efficiency of a Datacenter. PUE is determined by dividing the amount of power entering a Datacenter by the power used to run the computer infrastructure within it. PUE is therefore expressed as a ratio, with overall efficiency improving as the quotient decreases toward 1. PUE was created by members of the Green Grid, an industry group focused on Datacenter energy efficiency. Datacenter infrastructure efficiency (DCiE) is the reciprocal of PUE and is expressed as a percentage that improves as it approaches 100%. This term is used less often in the literature. Progress has been made towards standardizing this definition so that organizations can't deflate their PUE values by claiming efficiencies realized by recirculating hot air. Most existing Datacenters have a PUE close to 2.

Queuing Systems

See Batch Schedulers.

RAM

Random Access Memory

RISC and CISC

Reduced Instruction Set Computing is a CPU design that maintains that simplified instructions can eventually provide higher performance since it can execute faster with a smaller instruction set. This approach contrasts to complex instruction set computing in that CISC generally allows you to program in a higher-level language, allowing easier programmability and portability of applications. The vast majority of our supercomputing is done on CISC architectures these days. But there is a trend towards RISC architectures.

SAN — Storage Area Network

A storage area network is a dedicated storage network that provides access to consolidated, block level storage. NAS (Network Attached Storage) is often used in place of this term with minor differences.

SATA

Serial Advanced Technology Attachment is a computer bus interface for connecting host bus adapters to mass storage devices such as hard disk drives

SDR, DDR, QDR

Single, Double and Quad Data Rates, usually applied to network speeds.

SIMD — Single Instruction Multiple Data

A term used to describe a technique at the core of parallel processing. It implies performing the same calculation or computation on multiple data elements.

SMP — Symmetric Multiprocessing

A feature in our CPUs for over a decade; it is through this capability that we can put more processing elements (PE) or cores on the same die size where we could only previously put a single CPU core. Nowadays it is not unusual to find 4 or 8 PEs on an enterprise level CPU.

SWAR — SIMD Within A Register

TCO

Total Cost of Ownership. This aggregates the Capital Expenditure (CAPEX) cost and the Variable Expenditure (VAREX) or alternately the Operating Expenditure (OPEX), usually represented by power use associated with operating a Datacenter.

TCP and TCP/IP

Transmission Control Protocol and Internet Protocol are two networking protocols used in HPC. They “travel” on top of the Ethernet standard.

TOP500.Org

The site that lists the top 500 HPC centers and supercomputers in the world by Peak Performance. This is the de facto standard of comparison between performance capabilities of various supercomputer designs. The turnover in the Top500 is very high. The composite effects of many Moore's Law following technologies makes for a very rapid “obsolescence” of HPC equipment. The current list has systems of over 2 PFLOPs at the top and 31 TFLOPs at the bottom.

UPS — Uninterruptible Power Supply

Von Neumann Bottleneck (also referred to as the Memory Wall)

The separation between the CPU and memory leads to limiting the throughput relative to the amount of memory. In most modern computers throughput is much smaller than the rate at which the CPU can work. This seriously limits the effect of processing speed when the CPU is required to perform minimal processing on large amounts of data. The CPU must go into wait states before the needed data is transferred to or from memory. There are a number of methodologies that take part of the edge off of this problem, notably prefetching and predictive branching. This allows us to fill as much memory per cycle as we can with data that we may not use but it will be there in case we do. MPP has allowed us to avoid this bottleneck problem for years but this problem has now become untenable. This is the Number 2 enemy of Green HPC. A successful resolution of this would go a long way to solving our power problems and allow us to continue on our historical growth curve.

WOL

Wake-on-LAN is an energy saving function used on networks.

Works Cited

Andersen, D. G. (2009, May). *FAWN: A fast array of wimpy nodes.* Retrieved April 29, 2011, from Mendeley: http://www.mendeley.com/research/fawn-a-fast-array-of-wimpy-nodes-1/

Azevedo, D., Cooley, J., Patterson, M., & Blackburn, M. (2011). *Data Center Efficiency Metrics: mPUE™, Partial PUE, ERE, DCcE.* Retrieved April 29, 2011, from The Green Grid: www.thegreengrid.org/~/media/TechForumPresentations2011/Data_Center_Efficiency_Metrics_2011.ashx

Barroso, L. A., & Hölzl, U. (2007, December). *The Case for Energy-Proportional Computing.* Retrieved April 29, 2011, from Google Research Archives: http://research.google.com/pubs/archive/33387.pdf

Belady, P. C. (2007, February 1). *In the data center, power and cooling costs more than the IT equipment it supports.* Retrieved April 29, 2011, from Electronics Cooling: http://www.electronics-cooling.com/2007/02/in-the-data-center-power-and-cooling-costs-more-than-the-it-equipment-it-supports/

Borkar, S. Y., Dubey, P., Kahn, K. C., Kuck, D. J., Mulder, H., Pawlowski, S. S., et al. (2005). *Platform 2015: Intel Processor and Platform Evolution for the Next Decade.* Retrieved April 29, 2011, from Intel: http://epic.hpi.uni-potsdam.de/pub/Home/TrendsAndConceptsII2010/HW_Trends_borkar_2015.pdf

Broadcom. (2001). *Energy Efficient Network.* Retrieved April 29, 2011, from Broadcom: http://www.broadcom.com/products/features/energy_efficient_network.php

Clark, M., La Plante, P., & Greenhill, L. (2011, March 2). *Using GPUs for Signal Correlation.* Retrieved April 29, 2011, from Drop Box: http://dl.dropbox.com/u/1234422/correlator_scigpu.pdf

Cong, J. (2008). *Customizable Domain-Specific Computing.* Retrieved April 29, 2011, from UCLA Center for Domain-Specific Computing: http://www.cs.uci.edu/bin/pdf/seminarseries2k9/Cong.pdf

Data Direct Networks. (2011). *Exascaler.* Retrieved April 29, 2011, from Data Direct Networks: http://datadirectnet.com/exascaler

DCM. (2009, January 9). *Measuring and publishing UPS efficiency – a dark art?* Retrieved April 29, 2011, from Data Center Management: http://www.datacentremanagement.com/articles-features-71/power/823-measuring-and-publishing-ups-efficiency--a-dark-art.html

Dept. of the Army, U. S. Army Materiel Command, RDECOM Acquisition Center. (2006, September 20). *FBO #1759.* Retrieved April 29, 2011, from FBO Daily: http://www.fbodaily.com/archive/2006/09-September/20-Sep-2006/FBO-01147061.htm

Dietz, H. G. (2010, November). *MOG (MIMD On GPU).* Retrieved April 29, 2011, from University of Kentuky: http://aggregate.org/MOG/

Dietz, H. (2000). *Inside the KLAT2 supercomputer: the flat neighborhood network & 3DNow!* Retrieved April 29, 2011, from ars technica: http://arstechnica.com/old/content/2000/04/klat2.ars/3

Dolz, M. F., Fernandez, J. C., Mayo, R., & Quintana-Orti, E. S. (2010, Jan 1). *Energy Saving Cluster Roll: Power Saving System for Clusters.* Retrieved April 29, 2011, from Arnet Miner: http://www.arnetminer.org:8080/viewpub.do?pid=2776613

Ellsworth, J. M. (2010, June 23). *Michael J. Ellsworth, Jr.* Retrieved April 29, 2011, from Illinois U.: http://gladiator.ncsa.illinois.edu/PDFs/datacenter/10/ellsworth.pdf

Energetics Incorporated. (2010, February). *DOE Data Center Power Use Efficiency Summary Report.* Retrieved April 29, 2011, from US Dept. of Energy: www1.eere.energy.gov/team/pdfs/datacenter_report0204.pdf

Evans, T. (2004). *The Different Types of Air Conditioning Equipment for IT Environments.* Retrieved April 29, 2011, from American Power Conversion: www.ptsdcs.com/whitepapers/46.pdf

Fanara, A., Abelson, J., Bailey, A., Crossman, K., Shudak, R., Sullivan, A., et al. (2007, August 2). *Report to*

Congress on Server and Data Center Energy Efficiency Public Law 109-431. Retrieved April 29, 2011, from U.S. Environmental Protection Agency ENERGY STAR Program: http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf

Google. (2011). *Data center best practices.* Retrieved April 29, 2011, from Google: http://www.google.com/corporate/datacenter/best-practices.html

Google. (2009, Spring). *Insights into Google's PUE Results.* Retrieved April 29, 2011, from Google Docs: https://docs.google.com/present/view?id=df28585q_55f22tt3hq&pli=1

Green500.org. (2010, November). *The Green500 List - November 2010.* Retrieved April 29, 2011, from The Green500 List: http://www.green500.org/lists/2010/11/top/list.php

Gustafson, J. (2009, August 31). *Future Clusters.* Retrieved April 29, 2011, from Cluster 2009: http://cluster2009.org/keynote1.php

Harris, R. (2008, December 8). *Many-cores hit the memory wall.* Retrieved April 29, 2011, from StorageMojo: http://storagemojo.com/2008/12/08/many-cores-hit-the-memory-wall/

Hemmert, S. (2010). Green HPC: From Nice to Necessity. *Computing in Science and Engineering, Vol. 12 No. 6* , 8-10.

HPC Challenge. (2010). *HPC Challenge Awards Competition at SC10.* Retrieved April 29, 2011, from HPC Challenge: http://www.hpcchallenge.org

Intel. (2010). *Moore's Law.* Retrieved April 29, 2011, from Intel Technology: http://www.intel.com/technology/mooreslaw/

Kalim, S., & Shal, J. (2006). *Power Efficiency Metrics for the top500.* Retrieved April 29, 2011, from Next Generation Networking Systems Laboratory: http://sierra.ece.ucdavis.edu/papers/publicpresentations/HP_Workshop_files/Shalf_Afternoon_NERSC_Power.pdf

LessWatts.org. (2011). *Saving Power on Intel Systems with Linux: Tips.* Retrieved April 29, 2011, from Less Watts: http://www.lesswatts.org/tips/

Lovins, A. B. (2008, December 3). *Profitable Solutions to Climate, Oil, and Proliferation.* Retrieved April 29, 2011, from Harvard University Center for the Environment: http://www.environment.harvard.edu/video/future_of_energy/lovins/lovins__profitable_solutions_to_climate_oil_and_proliferation.pdf

Michaels. (2009, October 26). *CFP Bi-Annual All-Members' Meeting.* Retrieved April 29, 2011, from MIT Communications Futures Program: http://cfp.mit.edu/events/oct09/CFP Fall09 Presentations/Michaels Str Grid CSAIL 102809.ppt

Miller, R. (2011, April 29). *Solar Power in Data Centers: No Longer A Novelty?* Retrieved April 30, 2011, from Datacenter Knowledge: http://www.datacenterknowledge.com/archives/2011/04/29/solar-in-data-centers-no-longer-a-novelty/

Mills, S. (2010). *A Renewable Energy Powered Data Center's Road to Cost Saving and Environmental Stewardship.* Retrieved April 29, 2011, from Illinois U: http://gladiator.ncsa.illinois.edu/PDFs/datacenter/10/mills.pdf

Newman, S., & Palmintier, B. (2008). *Systems Thinking for Radically Efficient and Profitable Data Centers.* Retrieved April 29, 2011, from Rocky Mountain Institute: http://www.rmi.org/rmi/Library%2FE08-06_SystemsThinkingDataCenters

Pakin, S., Lang, M., & Kerbyso, D. J. (2009). *The reverse-acceleration model for programming petascale hybrid systems.* Retrieved April 29, 2011, from IBM: http://www.ccs3.lanl.gov/pal/publications/papers/Pakin2009:rev-accel.pdf

Park, J. (2011, April 7). *Data Center v1.0.* Retrieved April 29, 2011, from Open Compute Project: http://opencompute.org/specs/Open_Compute_Project_Data_Center_v1.0.pdf

Park, J. (2011, April 14). *Designing a Very Efficient Data Center.* Retrieved April 29, 2011, from Facebook Engineering: http://www.facebook.com/notes/facebook-engineering/designing-a-very-efficient-data-center/10150148003778920

Ramanunni, J. (2011, April 29). *Go Modular With HP's New Data Centres.* Retrieved April 30, 2011, from EFY Times: http://www.efytimes.com/e1/fullnews.asp?edid=62205

Rasmussen, N. (2003). *Avoidable Mistakes that Compromise Cooling Performance in Data Centers and Network Rooms.* Retrieved April 29, 2011, from American Power Conversion: http://www.ptsdcs.com/whitepapers/42.pdf

Strzodka, R., & Goddeke, D. (2008). *Mixed Precision Methods on GPUs.* Retrieved April 29, 2011, from NVIDIA: http://www.nvidia.com/content/nvision2008/tech_presentations/NVIDIA_Research_Summit/NVISION08-Mixed_Precision_Methods_on_GPUs.pdf

Top500.org. (2010, November). *Interconnect Family share for 11/2010.* Retrieved April 29, 2011, from Top 500 Computer Sites: http://www.top500.org/stats/list/36/connfam

Tschudi, W., Proggens, P., Bell, G., Mathew, P., Sartor, D., & Pfeifer, R. (2009, January 1). *Proposed LEED Criteria for Data Centers.* Retrieved April 29, 2011, from Berkeley Lab: http://hightech.lbl.gov/dc-epc.html

U.S. Dept of Energy. (2010). *Purchasing More Energy-Efficient Servers, UPSs, and PDUs.* Retrieved April 29, 2011, from Energy Star: http://www.energystar.gov/index.cfm?c=power_mgt.datacenter_efficiency_purchasing

von Laszewski, G., Wang, L., Younge, A., & He, X. (2009, August 31). *Power-Aware Scheduling of Virtual Machines in DVFS-enabled Clusters.* Retrieved April 29, 2011, from IEEE Cluster 2009: http://cluster2009.org/paper16.php