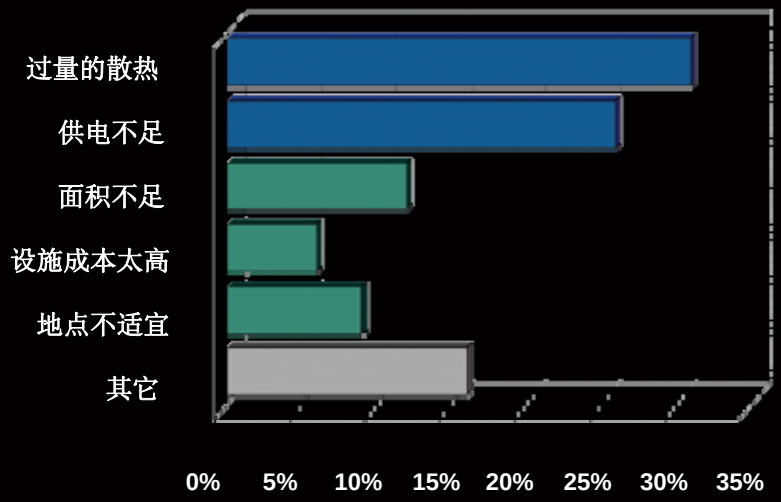
李丽芳
资深系统架构师
lilf@cn.ibm.com
IBM系统与科技事业部

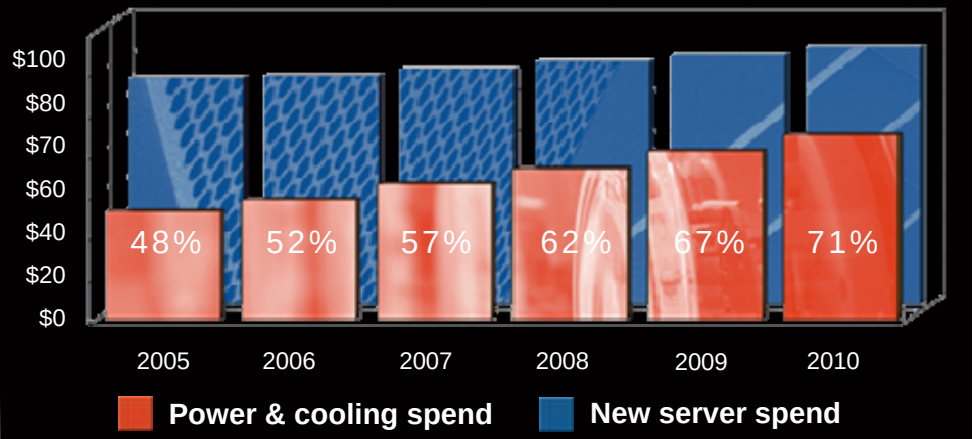
为什么需要绿色数据中心?

- 高耗电密度和迅速增长
- 每平方米能源消耗10倍甚至100倍于普通办公大楼
- 对供电系统的潜在影响巨大
- 2005年美国数据中心消耗约为450亿度电
- 以目前的发展速度，能源需求可能在5年内倍增

数据中心最大的设施问题是什么? Gartner 2006



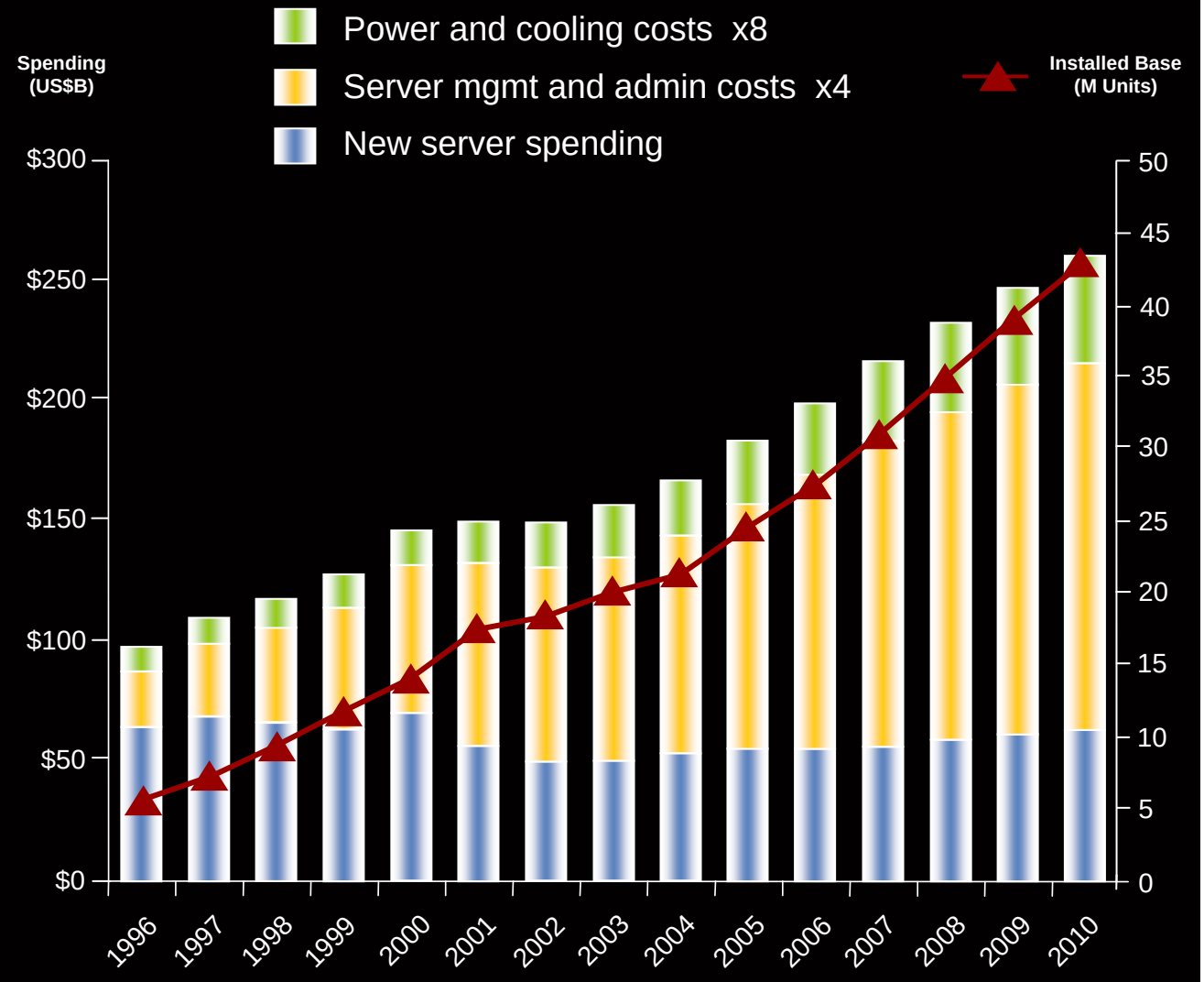
服务器安装的供电和冷却成本 ($US)



全世界IT 开销趋势

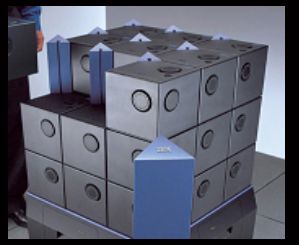
今天，在硬件上花一块钱，意味着要在能源上花**5**角钱.

而在未来**4**年，这个比例会继续增长**54%.**



Source: IDC, Virtualization 2.0: The Next Phase in Customer Adoption, Doc #204904, Dec 2006

IBM 2007年专利数量: 连续15年领先于其它公司



绿色专利
系统与技术部: **350+**
IBM: 1300+

IBM 名列绿色IT制造商第一名.....

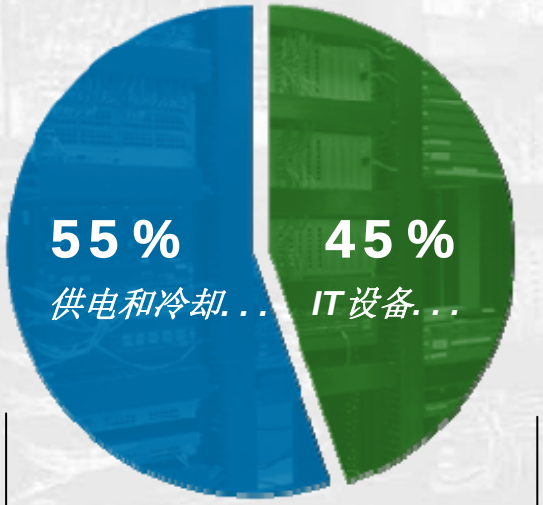
- 2008年2月15日，”计算机世界”
- 绿色计算是**IT**和环境的巨大胜利：它提供了一个省钱的机会。许多公司正在努力使**IT**变得更加节能，但是只有很少的公司能真正做到。下面是前**12**名**IT**制造商：
- ***1 . IBM***
- ***6.*** *Microsoft Corp.*
- ***7.*** *Hewlett-Packard Co.*
- ***12.*** *Sun Microsystems Inc.*

IBM client, **Highmark**, was named Computerworld's top Green-IT User after implementing energy-saving technologies, such as server virtualization

http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=312485

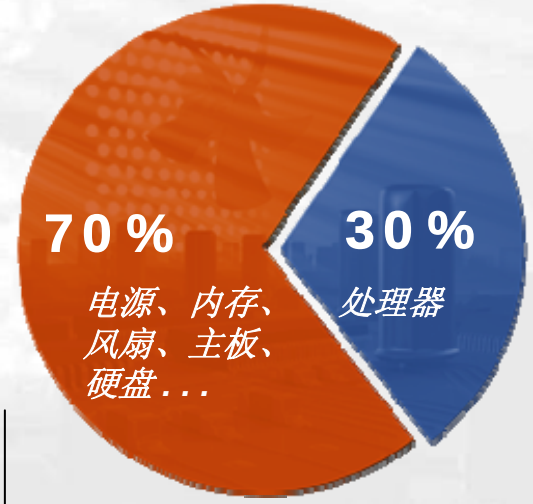
数据中心提高能效之道

数据中心



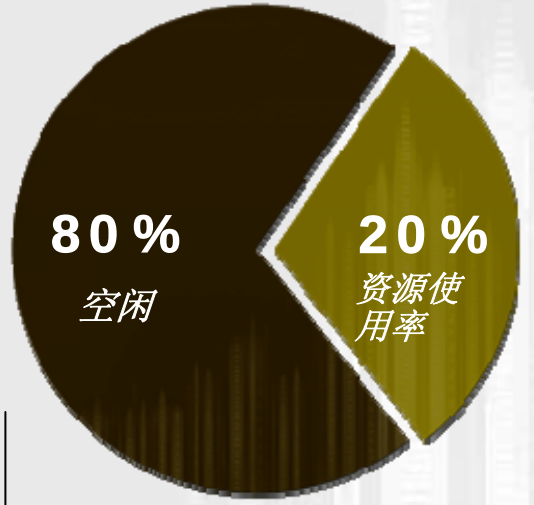
- 监测并智能调整服务器的用电
- 降低数据中心整体耗电
- 提高供电和冷却的能源效率
- 提高IT设备用电在数据中心总用电的比例

服务器硬件



- 在提高CPU性能的同时降低或保持能耗水平
- 提高服务器电源的能源效率
- 提高服务器送风冷却的能源效率

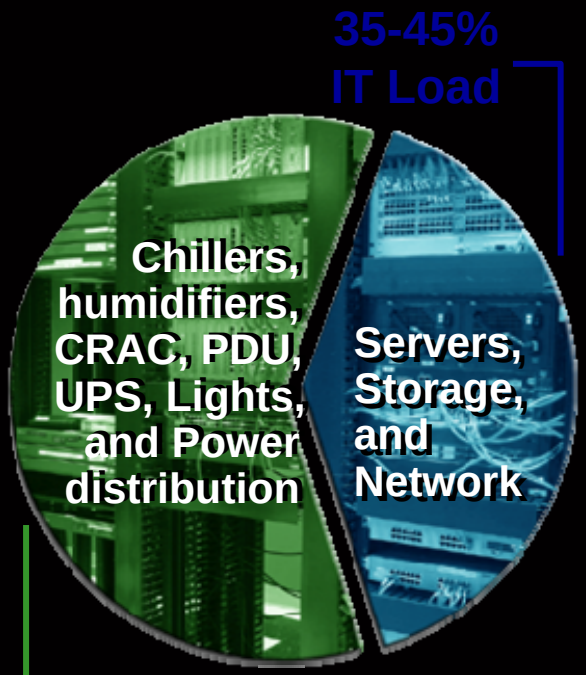
服务器负荷



- 在保证响应时间和服务水平的前提下，提高服务器的使用率

数据中心能耗分析和提高能效之道

Data Center

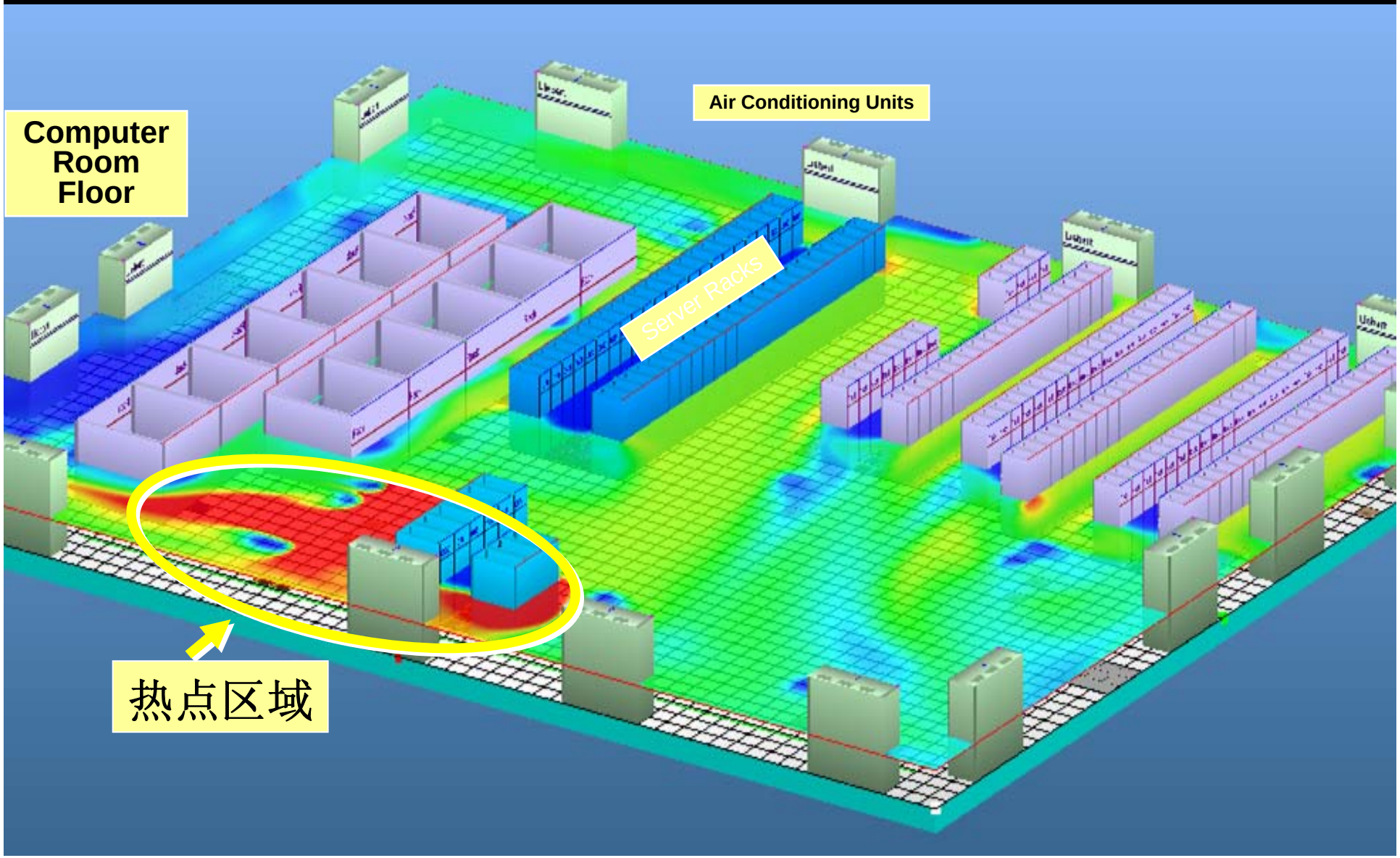


Data source: *Creating Energy-Efficient Data Centers, , U.S. Department of Energy , Data Center Facilities and Engineering Conference , May 18, 2007*

IBM 为客户提供的服务…

- ✓Data Center Stored Cooling Solution
- ✓Optimized Airflow Assessment for Cabling
- ✓Scalable Modular Data Center
- ✓Data Center Relocation and Consolidation, Data Center Facilities Design
- ✓Data Center Energy Efficiency Assessment
- ✓Data Center Thermal Analysis and Optimization Facilities Integration
- ✓IBM Optimization and Integration Services: Server Consolidation
- ✓Accelerator for Rationalization

IBM 技术帮助减少数据中心的热点



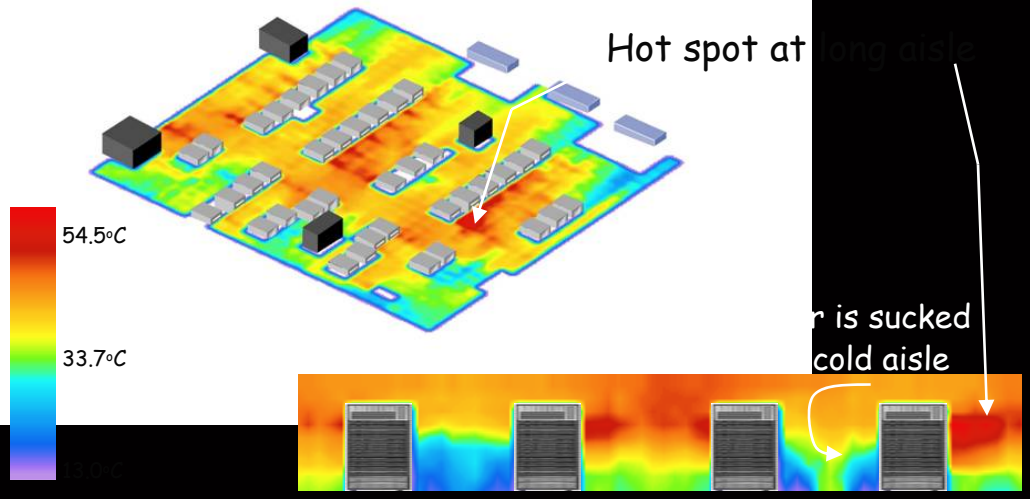
移动测量技术(MMT)

Thermal Analysis and Optimization - Pilot Phase

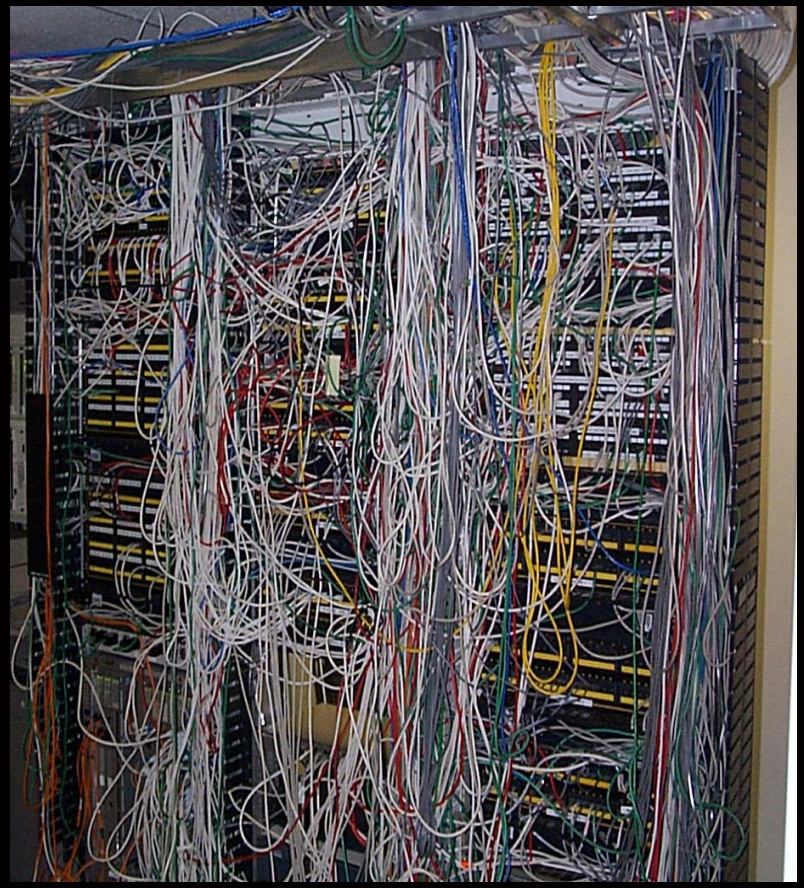
Optimize data center thermal profile to eliminate hot spots and reduce energy consumption.

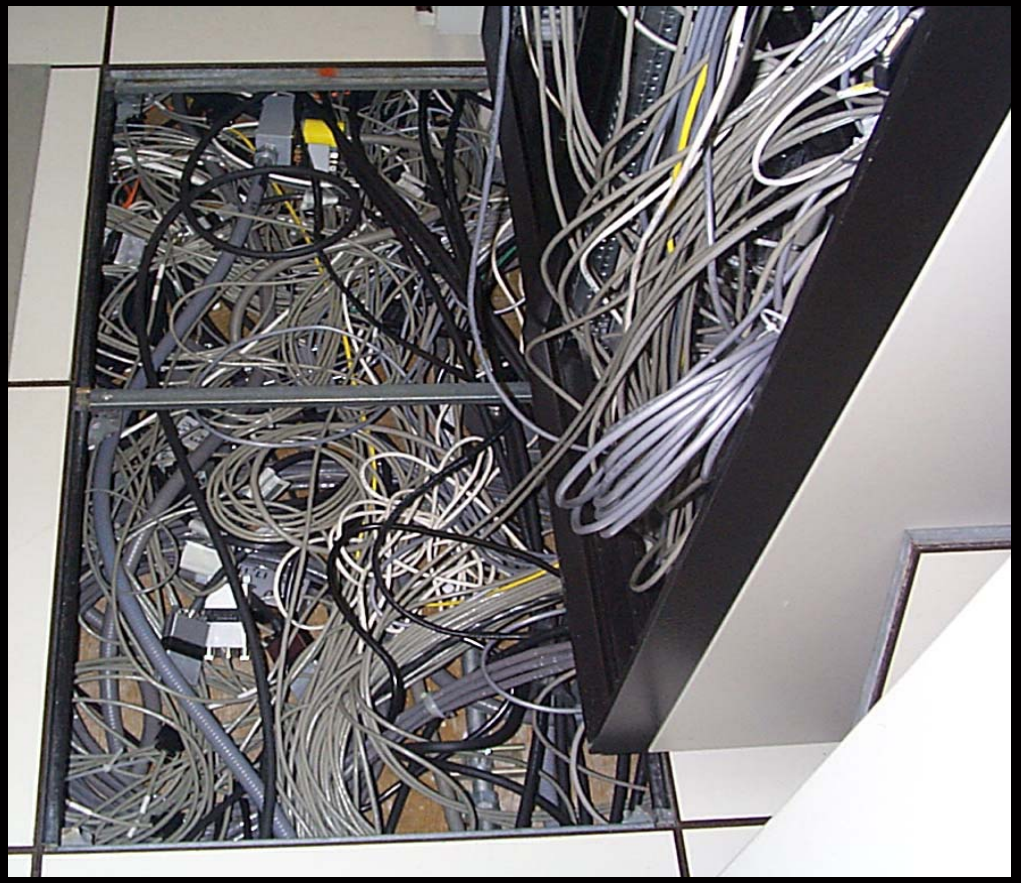
Utilizing MMT

- Unique cart-based design for measurement collection
- Creates a 3D temperature map of the data center
- Samples 100+ temperatures at one time and thousands of temperatures in the data center
- Monitors position
- Surveys large areas in a short time



混乱的布线.....





针对布线的气流优化评估服务

Under-floor savings

- Description:
- **A comprehensive assessment of the existing data center cabling infrastructure provides an expert analysis of the overall cabling actions. The service is designed to improve airflow for optimized cooling, simplified change management and improved cabling resiliency.**

Before

- Potential benefits:
- **Improved airflow under raised floor: helps create a more energy-efficient data center**
- **Fewer “hot spots” due to bypass airflow**
- **Improved manageability for all cabling systems**
- **Savings on operational costs—reductions associated with cable installation and change management**
- **Improved resiliency—elimination of potential for points of failure**

After

数据中心能耗分析和提高能效之道

Data Center

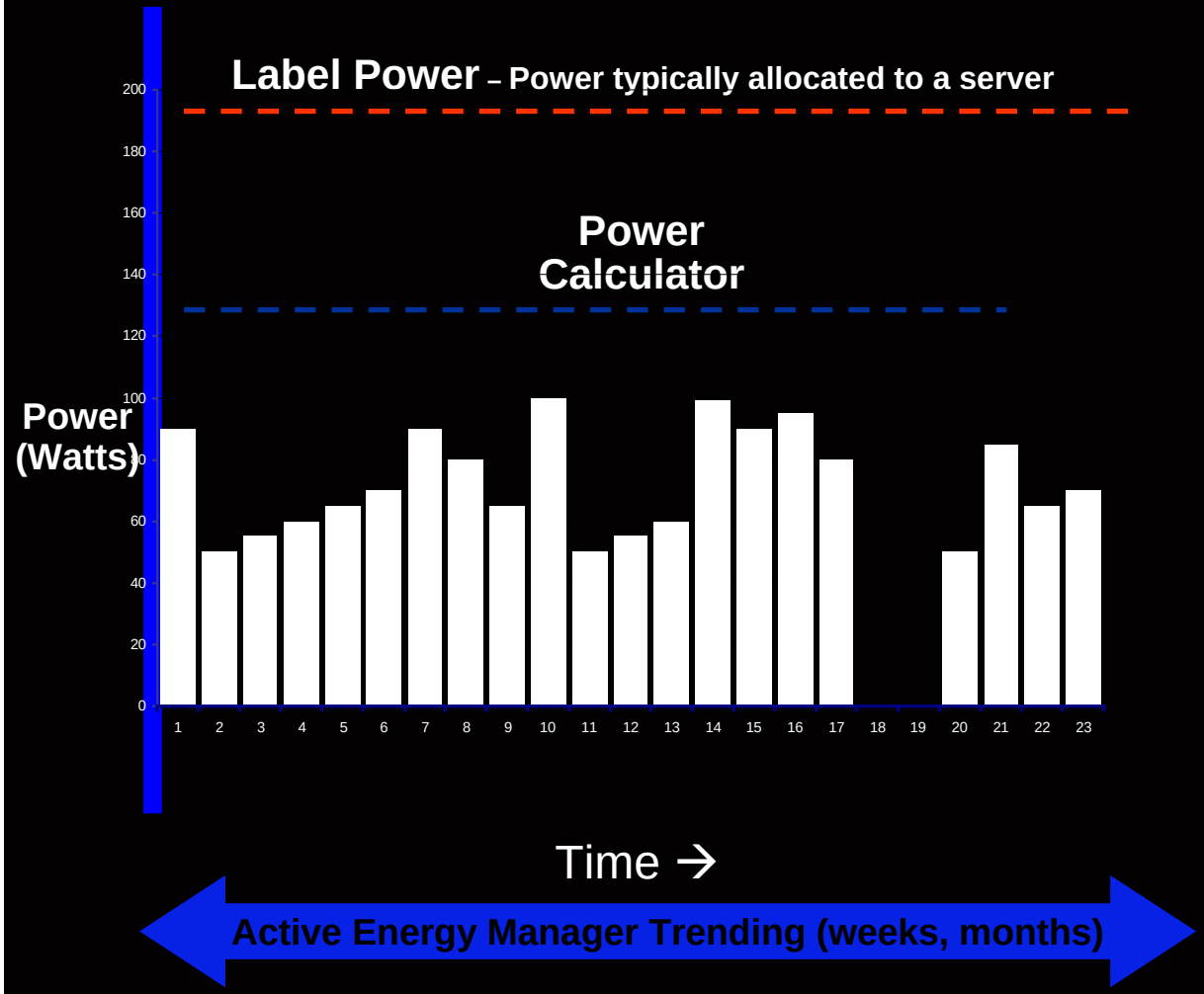
Server Hardware

IBM Power Systems的绿色能源技术

- IBM的又一创新: **EnergyScale™**
  - 综合的硬件和系统管理工具
  - 让客户能够优化能源的消耗，管理和冷却
    - 在服务器、机架和数据中心三个层面.
- 主要组成:
  - 系统计算器 – 规划服务器和电源配置的估算工具
  - 背门热交换器 – 机柜的水冷解决方案
  - **System p** 的技术参数 – 对核心系统部件设定耗电上限
  - **Active Energy Manager** – 管理能耗的先进工具，提供报告、配置、控制等的功能
  - 能耗和虚拟化的集成 – 优化和管理工作负载，尽量减少能耗
  - **POWER6 Nap** 模式
  - **EnergyScale I/O**
  - **Power Saver Mode**
  - **Power Capping**

细化到服务器水平的电力分配

Manage furnished peak power to actual used



Power Allocation Model to Server

Wasted Power Allocation

Power Budget Not Converted into Compute Cycles

Proper Power Allocation

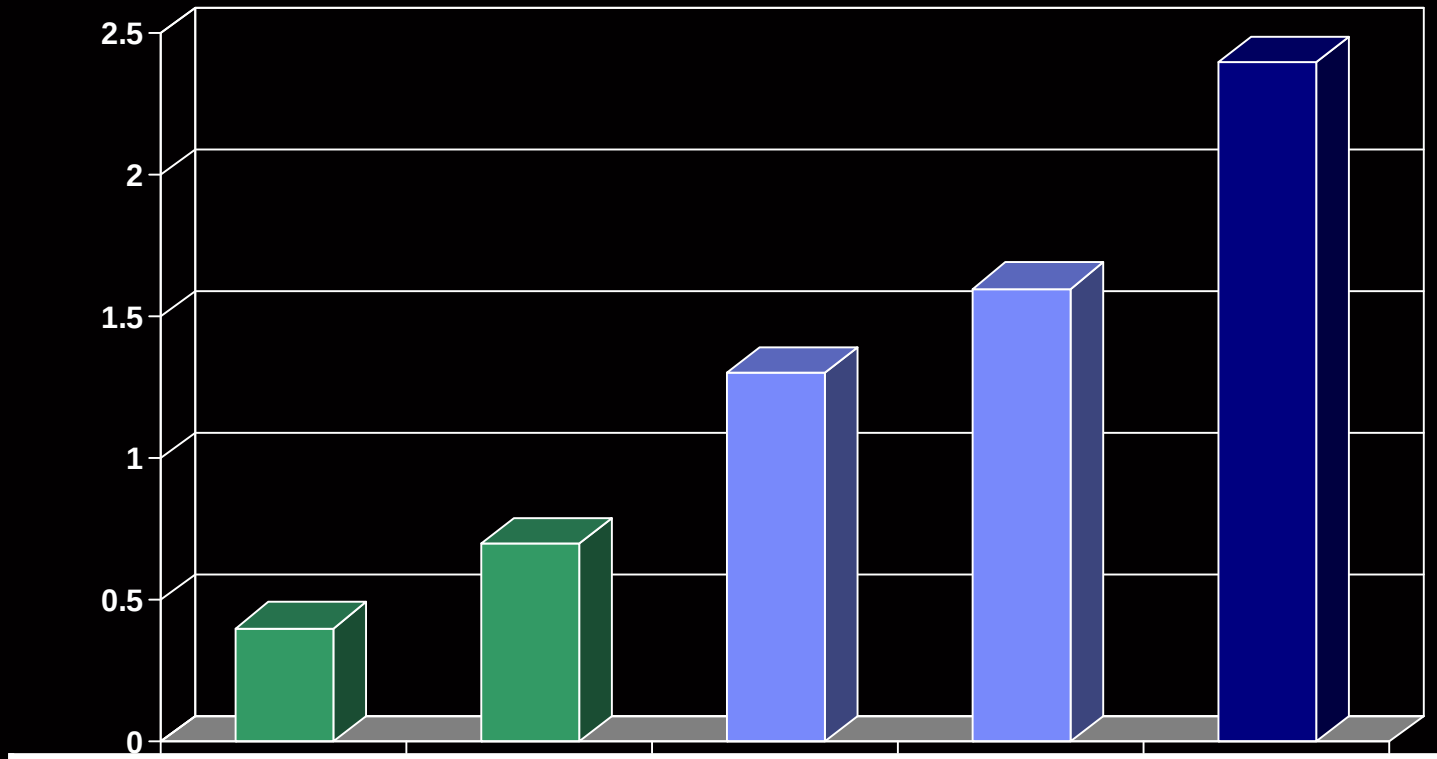
Power Budget Converted into Compute Cycles

系统计算器(System Calculators)

- 电源负荷计算器
  - Obtain an estimate of the typical heat output (watts and BTU/hr)
  - Power source loading (kVA) for a specific configuration under normal operating conditions.
  - Website: http://publib.boulder.ibm.com/infocenter/eserver/v1r3s/index.jsp?topic=/iphdl/systemcalculators.htm
- 地板承重计算器
  - Determines the distributed load that a server, expansion unit, or migration tower places on a subfloor (i.e. a raised floor) or on an above ground, non-raised floor structure.

性能提升的同时减少能源的消耗

rPerf per KWatt



POWER4™ p670 1.1 GHz rPerf: 24.46 KWatts: 6.71	POWER4+™ p670 1.5 GHz rPerf: 46.79 KWatts: 6.71	POWER5™ p5-570 1.65 GHz rPerf: 68.4 KWatts: 5.2	POWER5+™ p570 1.9 GHz rPerf: 85.20 KWatts: 5.2	POWER6™ p570 4.7 GHz rPerf: 134.35 KWatts: 5.6

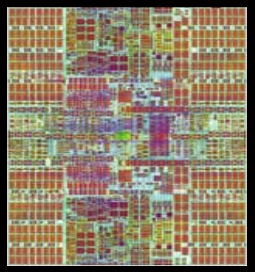
IBM Power6^{T} 服务器 — 提高能效

- POWER6 芯片的功能
  - **Power Reduction:** Monitor & reduce power to idle logic within cores
  - **NAP Mode:** Power off inactive cores, restore power when needed
  - **Thermal Tuning:** Sensors monitor & reduce power to overactive circuits
  - **Virtualization:** Moving running UNIX and Linux operating system workloads from one POWER6 server to another.

- POWER6 服务器的功能
  - **Enhanced System Design & Implementation:** Improved server Performance / WATT uplift over POWER5*.
  - **EnergyScale I/O:** Powering off PCI slots not being used
  - **Variable Fan Speed (10,500 – 5500 RPM):** Reduces power to fans (1/3 of total server power) by up to 45% based on ambient temperature**
  - **Rear Door Heat Exchanger:** Cools exhaust air from 19 & 24” rack, removes up to 60% of the heat#

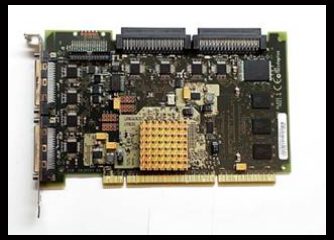
* Based on various SPEC benchmarks; IBM p570 POWER6 result to be submitted on 5/21/07
** Based on IBM internal measurements
# IBM press release; 05/10/07; http://www.ibm.com/press/us/en/pressrelease/21517.wss

节能技术: POWER6 Nap 模式



- Problem: No active software thread to run on a processor
- Solution: POWER6 Nap Mode. Each hardware thread on a given processor core can issue an instruction putting itself into Nap mode. If both hardware threads for that core are in Nap mode, the whole processor core then enters the Nap state.
  - Nap state: The core eliminates almost all switching power
  - Nap Mode: Approximately 11% power savings over running idle loops
  - POWER6 cores enter and exit Nap independently of each other.
- Processor Nap state is interruptible; Operating system or hypervisor can re-awaken a napping core
- Nap Mode reduces the power consumed by *Capacity on Demand* cores:
  - POWER4 / 5: CoD cores consumed full power & ran an idle loop
  - POWER6 systems, all unlicensed cores are kept in Nap state

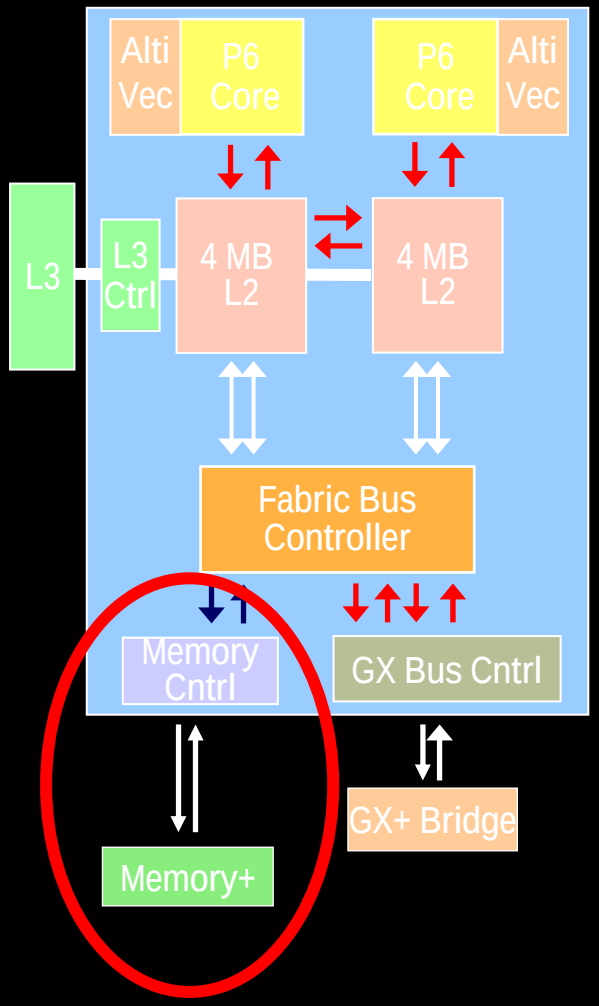
节能技术: EnergyScale I/O



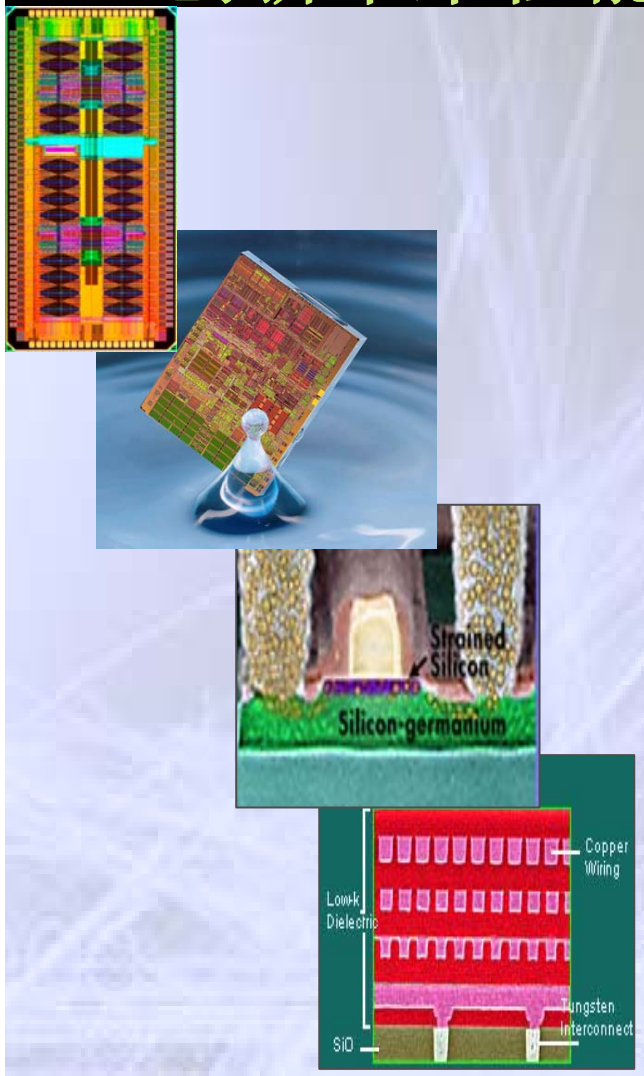
- Problem: Empty / Unused PCI slots drawing power
  - All PCI slots drawing power regardless of status.
- Solution: Automatically power off hot-pluggable PCI adapter slots that are not being used
  - Empty Slots (No adapters present)
  - Unused Slots (Slots not assigned to a partition)
  - Slots belonging to a partition not powered on
  - Applies to hot pluggable PCI slots only
  - Save up to 14 watts per slot
  - PCI slot is powered off immediately by system firmware when it is dynamically removed from the partition / powered off / etc..
- Support is available for all POWER6 processor based System i and System p servers, and the expansion units that they support
- System firmware automatically scans all hot pluggable PCI slots at regular intervals looking for ones that meet the criteria for being not in use and powers them off.

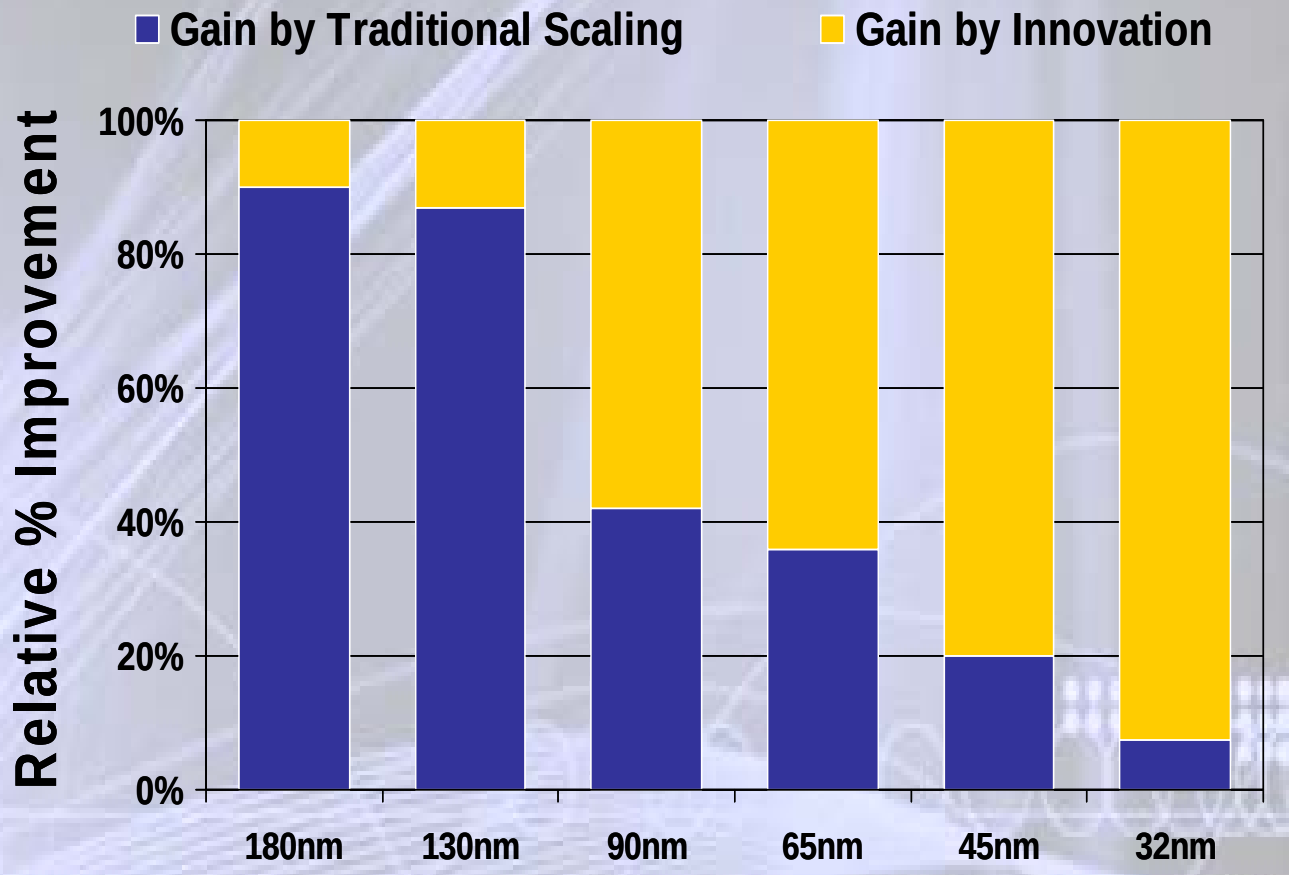
节能技术：内存控制器动态模式

- Problem: Memory subsystem consumes a significant fraction of the power Servers typically have large amounts memory
  - Large portion of the DRAM power is consumed by *idle* chips
  - Workloads cannot keep all the DRAM chips constantly busy.
- Solution: Utilize *Memory Powerdown technology*.
  - DRAM manufacturers provide a Lower-power idle mode, *powerdown*¸ to decrease the DRAM idle power by deactivating the clock-enable control signal to the DRAM chips.
- POWER6 on chip memory controller, exploits the DRAM *powerdown*
  - Enables significant savings in the memory subsystem power.
  - DRAMs are removed from *powerdown* as soon as a request to that rank is queued in the controller or when the rank must be *refreshed.*.
  - *Queue-driven* policy implemented in the POWER6 controller rarely see loss of performance due to this overhead, yet the system can obtain significant reduction in DRAM power consumption using this mechanism.



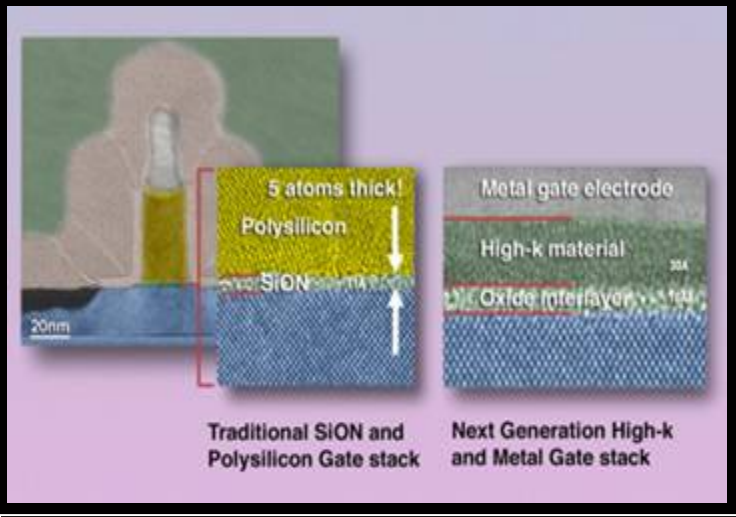
创新带来性能的飞跃





节能技术: High-K 技术

- Problem: How to Improve performance without excessive power & leakage?
  - Performance improvement achieved through “scaling”- shrinking chip dimensions
  - Gate leakage becoming much worse as chips are scaled
  - Drives excessive power and leakage
- Solution: High-K Metal Gates
  - New Substrate material: Hafnium
  - Much improved gate leakage
- Benefit: Better processors
  - Smaller chips
  - Faster processors with higher performance
  - Lower leakage and power



IBM High-k technology allows transistors to continue shrinking while using less power and improving performance.

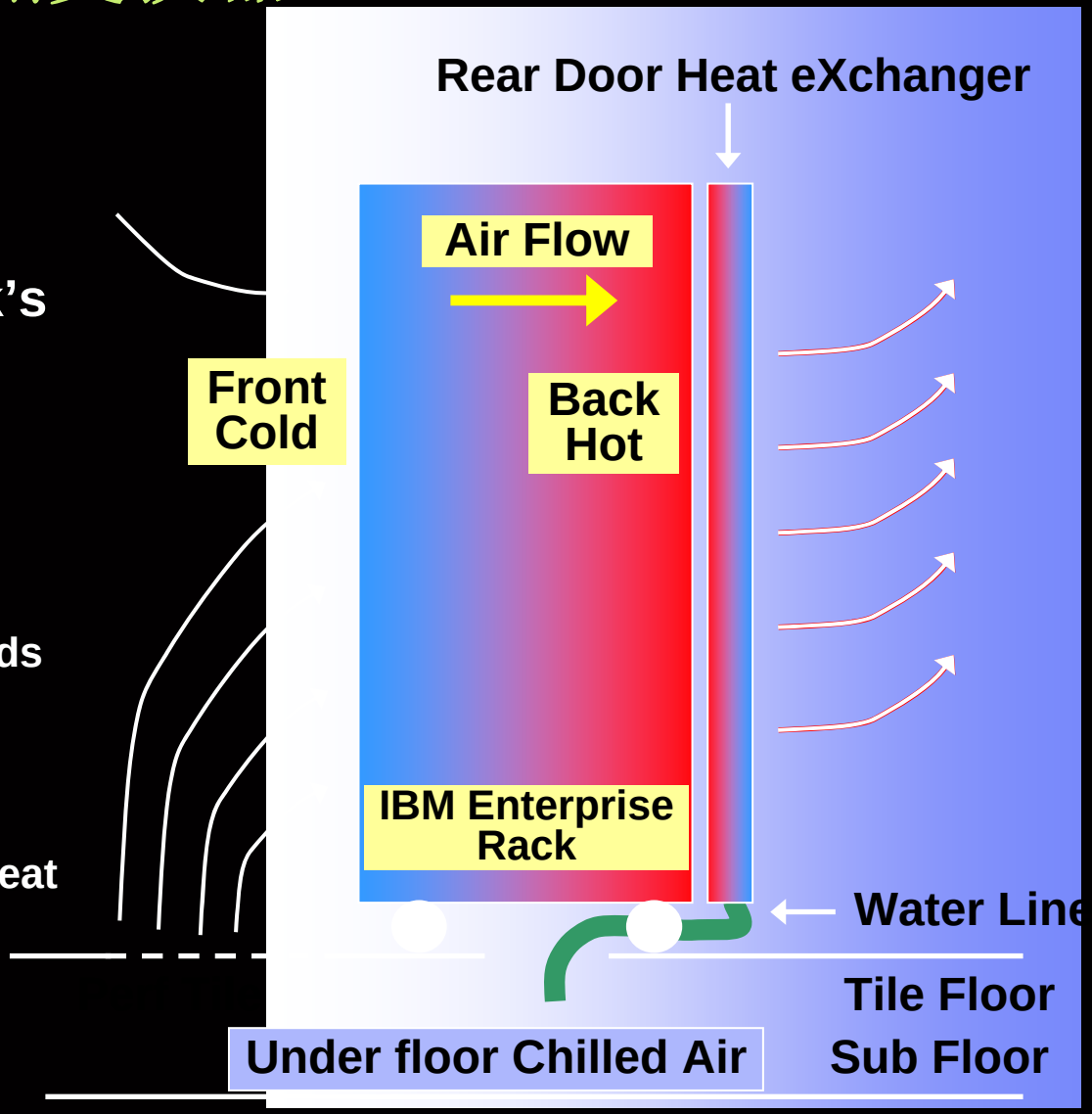
节能技术：背门热交换器

IBM’s CoolBlue Initiative includes Rear Door Heat eXchanger which can remove over 50% of a rack’s heat output

- No new fans or electrical load.
- Attaches to back of rack (adds 5”)
  No rearrangement of datacenter

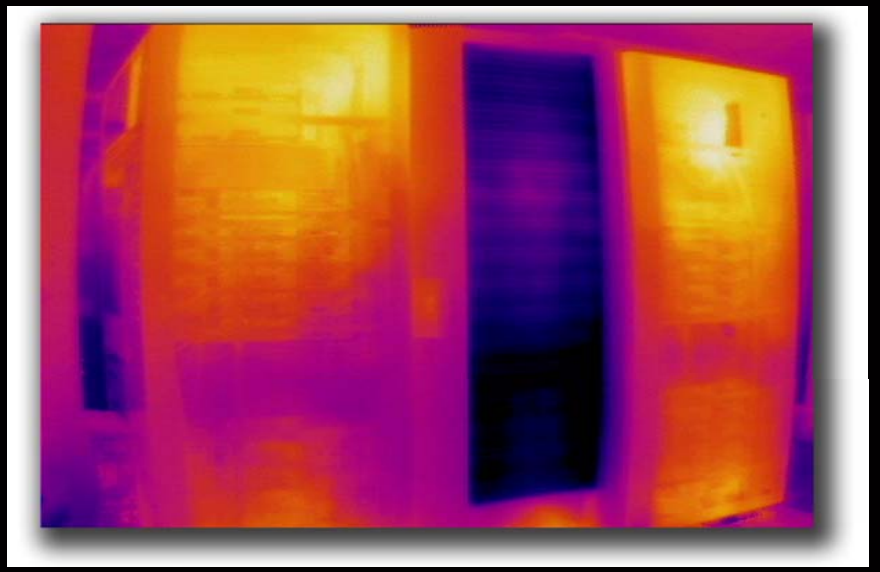
- Rear Door Heat eXchanger adds cooling capacity at ~1/4 of the cost of traditional methods

- Water can carry 3500x more heat than air at sea level



背门热交换器实景图

Visual images of the
Rear Door Heat eXchanger



Thermal images of the
Rear Door Heat eXchanger

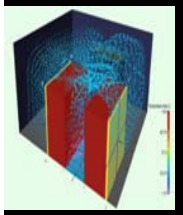
Removes over 50% of a
rack’s heat output

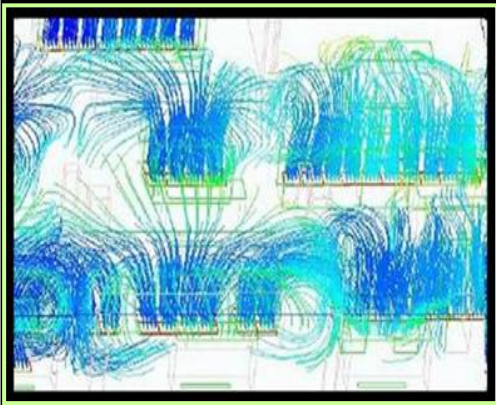
PG&E节能成功案例

PG&E Corporation

- 在4,000平方米的数据中心减少80%的能源消耗
- 合作研发热力建模技术——Mobile Measurement Technology
- 单个系统使用率得到10-80%的提高
  - 合并300台UNIX服务器至 6台IBM System p5™服务器上
- 通过IBM技术实现高达60%的散热减少
  - 实施IBM背门热换器直接从散热源带的60%散热

IBM Heat eXchanger





IBM Mobile Measurement Technology

IBM 能耗管理——测量、预测、封顶

Active Energy Manager

- 帮助用户测量和控制IT设备的用电
- 提供“Cruise control” 功能
- 必要时对IT设备用电进行封顶
- *适用于IBM所有的服务器和存储系统中*

Active Energy Management

Manage Power at the rack and server level

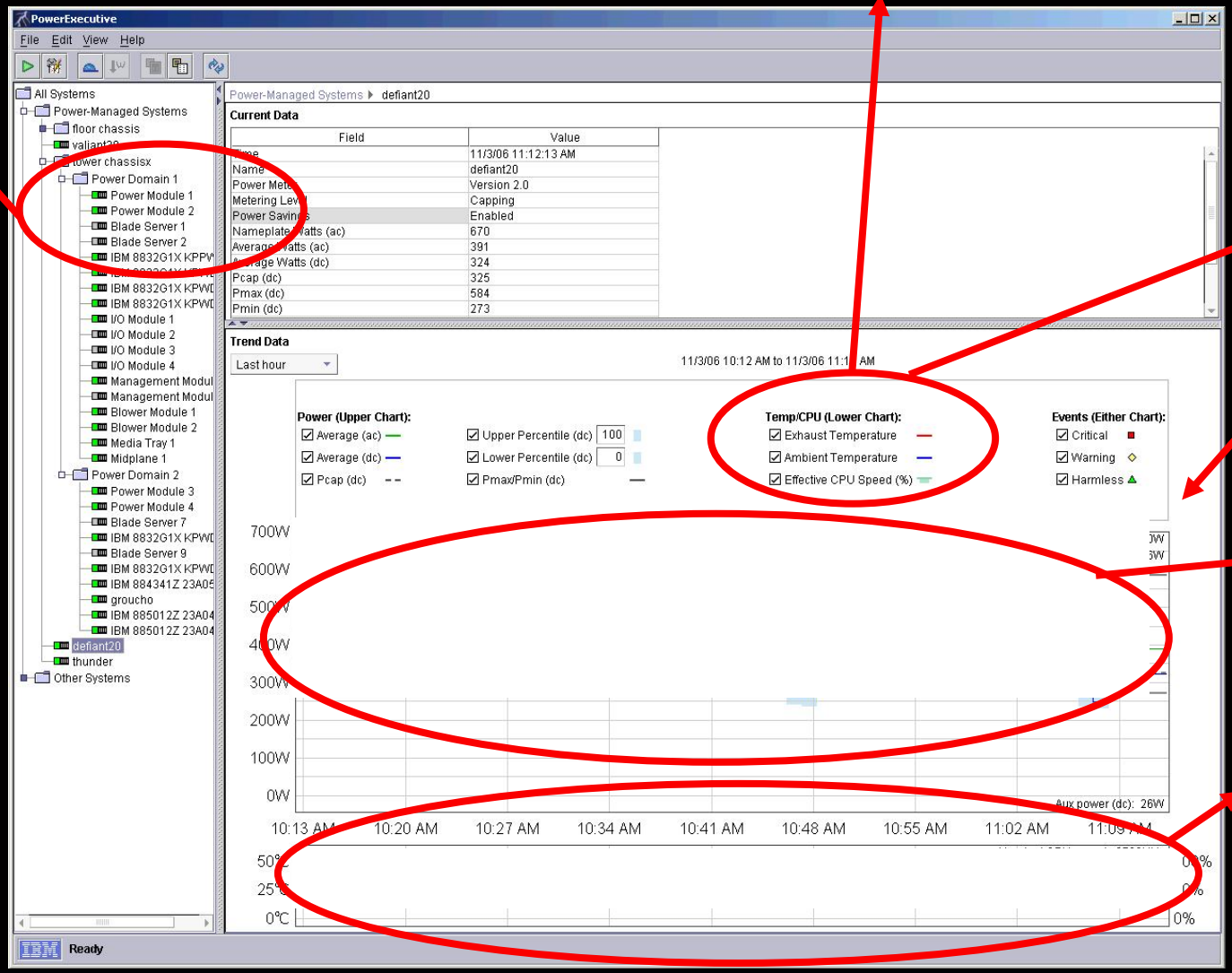
View inlet and exhaust temperature

Track heat emitted

Compare rack actual power vs. Label Power

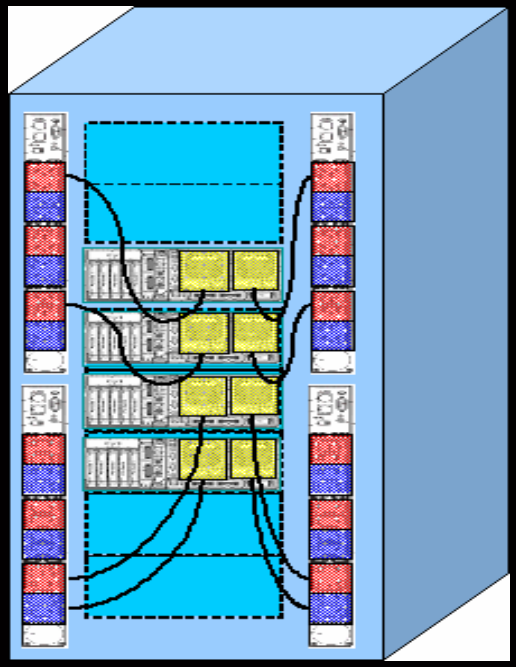
Trend power use over time

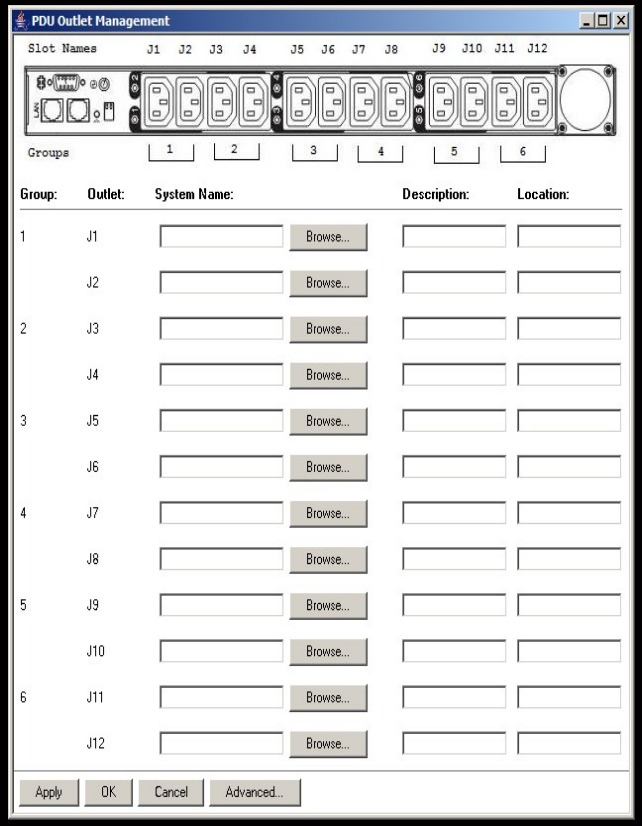
Trend temperature over time



Active Energy Manager – 支持iPDU

- Discover and monitor iPDUs:
  - Ships with racks
  - Display trending information per load group
- Allows mgmt of POWER6 p570, p5-570, Legacy & non-IBM Storage, and other non-server IT equipment
- Associate Director Managed Objects with load groups





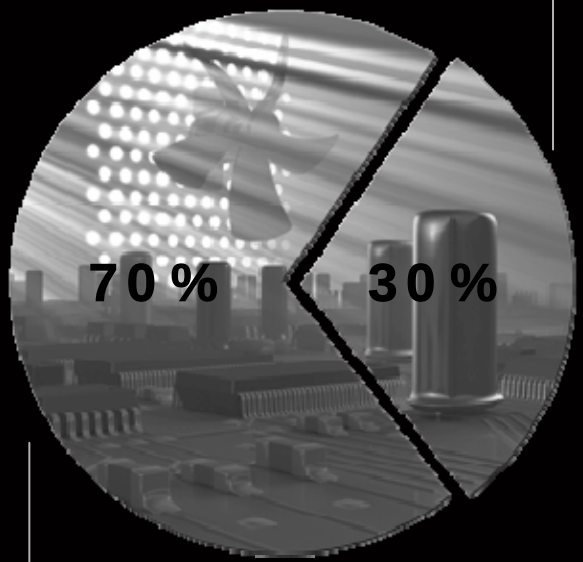
iPDU = Intelligent Power Distribution Unit

PowerExec sums power from both PDUs to show total server power

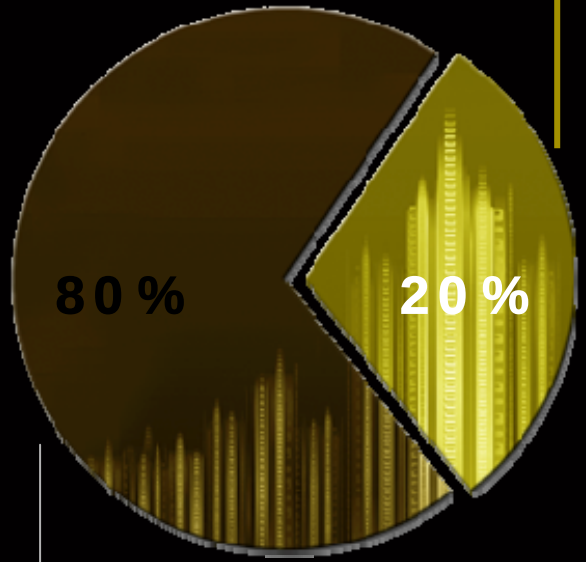
数据中心能耗分析和提高能效之道

Data Center

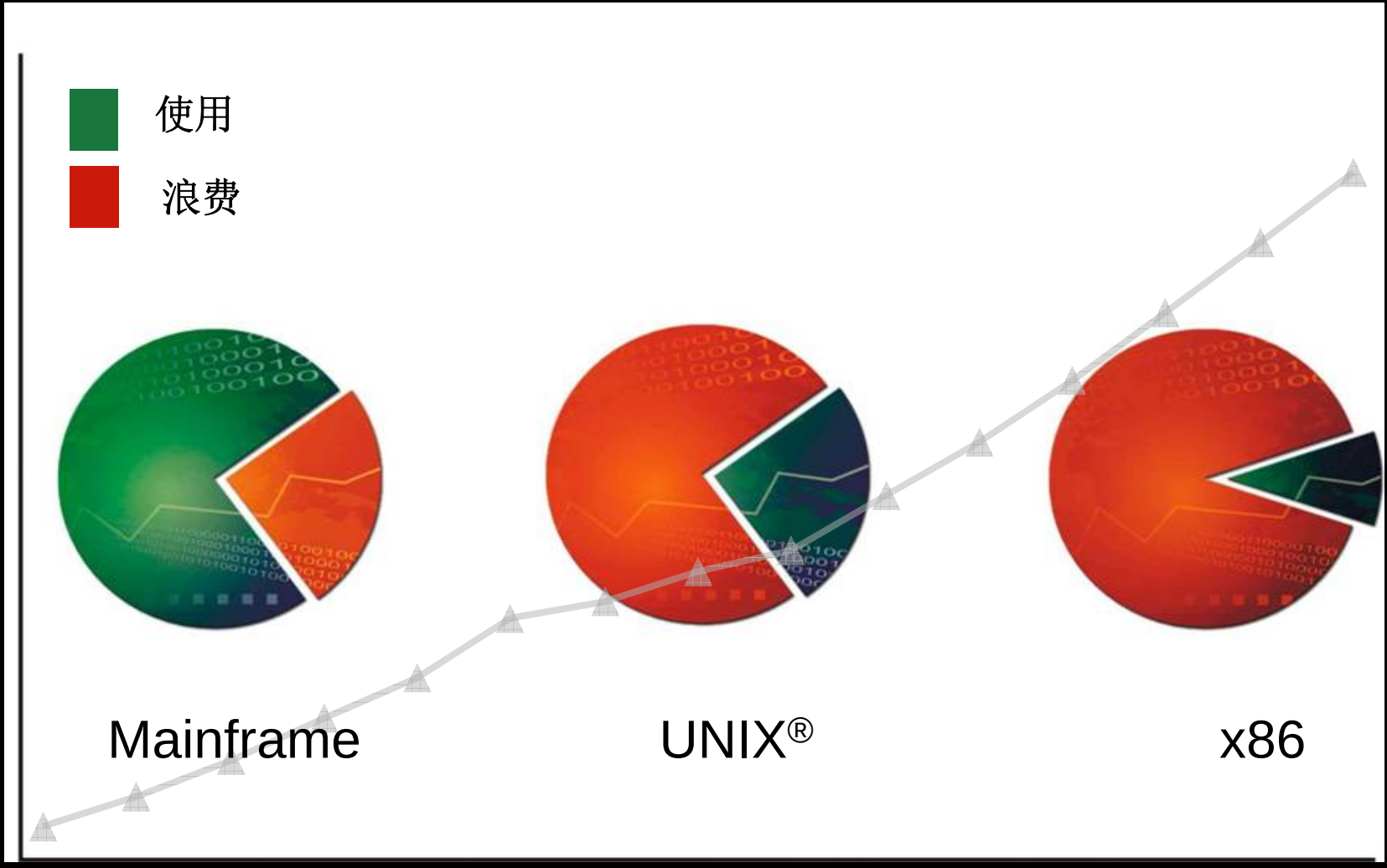
Server Hardware



Server Loads

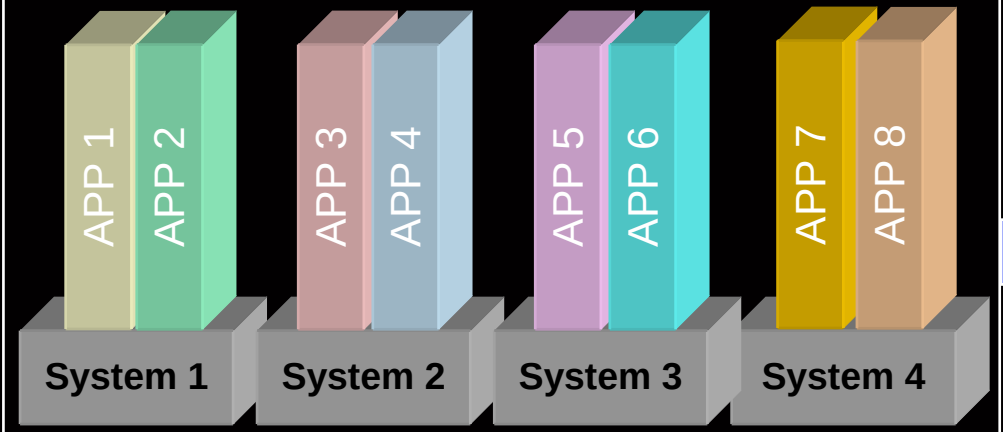


不同类型服务器的使用率



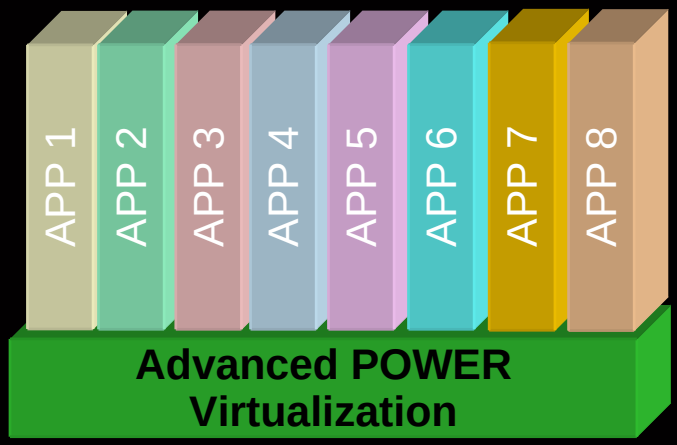
服务器合并提高服务器的使用率

Data Center Capacity 12KW, 67% used



10% busy 2KW | 10% busy 2KW | 10% busy 2KW | 10% busy 2KW

Data Center Capacity 8KW, 50% used



70% busy 4KW

使用虚拟化技术进行服务器合并是一个非常有效的节约能源的工具

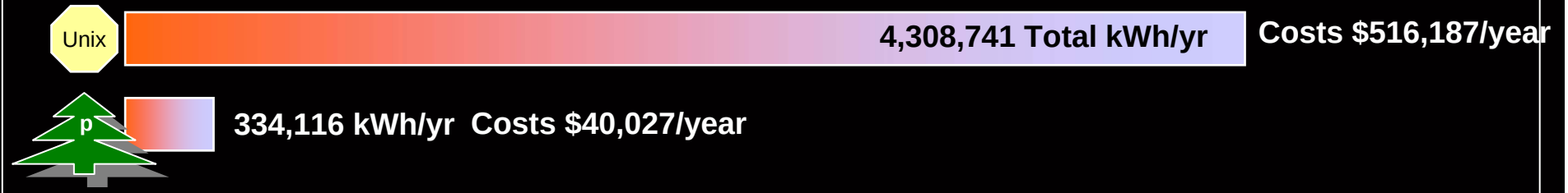
System p虚拟化方案的节能效果

用IBM先进POWER虚拟化替代传统Unix主机

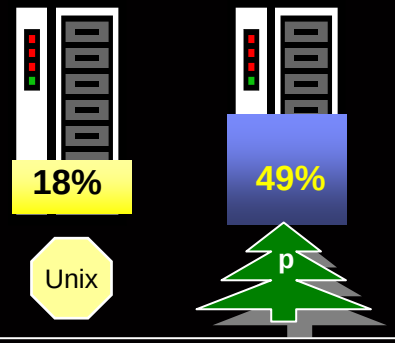
负荷类型

2% each for e-mail & collaboration & File/Print　4% for System Management 5% for Web

能耗比较



服务器使用率



服务器数量和占用空间

752台物理服务器 + 95个机柜

7台物理服务器/36个逻辑分区 + 3个机柜

Source: based on large studies conducted by IBM Lab Services. The analysis indicates projected energy savings which represent averages of more than 4,000 servers in actual customer data centers. Actual customer data is used for competitive systems.

IBM POWER6领先的虚拟化技术和功能

IBM PowerVM 技术

分区负荷管理器
- 自动平衡对处理器和内存的请求

集成虚拟化管理器
- 不需要HMC（硬件控制台）便可管理逻辑分区

虚拟I/O服务器
- 简化以太网、SCSI和光纤通道连接

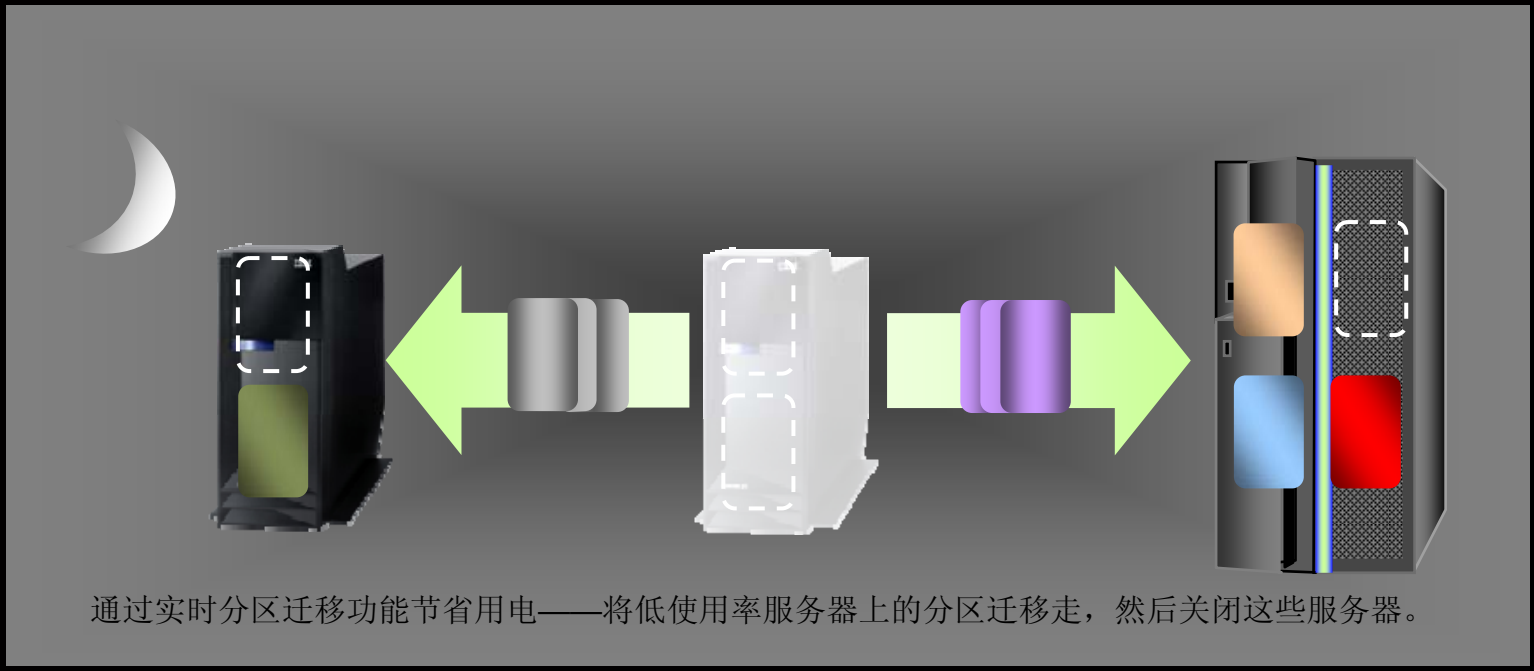
实时分区迁移和实时应用迁移
- 将逻辑分区在P6服务器之间动态迁移，不中断运行
- 将负载分区在P6服务器之间动态迁移，不中断运行

Micro-Partitioning™ (POWER Hypervisor)
- 单个处理器可用于创建多至10个的微分区——单台服务器可创建多至254个
- 动态调整分区的大小（处理、内存等）而不中断运行
- 通过类似System z的PR/SM硬件微码实现——精简、优化、整合

- 业界领先的虚拟化技术，支持**AIX**和**Linux**操作系统
- 实现在硬件实用率和灵活性的显著提高
- 从**2004**年**8**月开始在全球用户中广泛实施

POWER6的虚拟化新功能——实时分区迁移

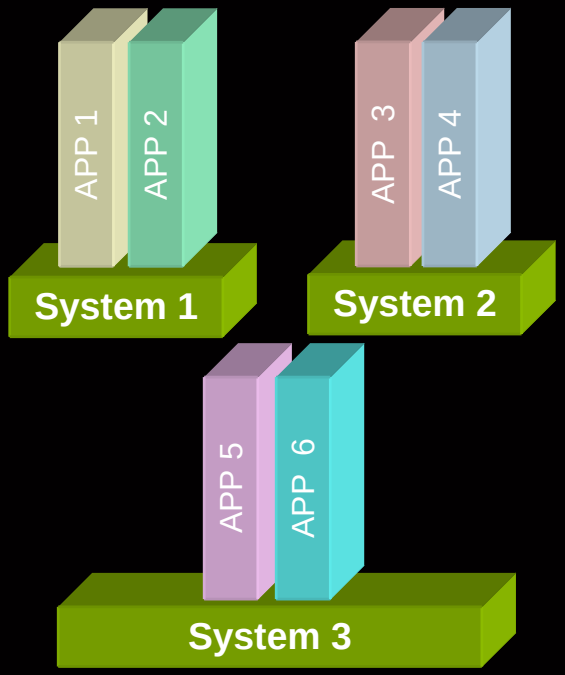
- *在非高峰时间，关闭部份数量的服务器*



通过实时分区迁移功能节省用电——将低使用率服务器上的分区迁移走，然后关闭这些服务器。

基于策略的虚拟化示意

虚拟化+工作负荷迁移实现 *动态的服务器合并*



把休眠、关机和其它低电状态与工作负荷动态平衡，以及自动部署工具结合起来可以形成非常有效的用电和散热管理系统

自动用电控制
基于策略的自动化

减少整体用电
转移合并工作负荷

虚拟化成功案例

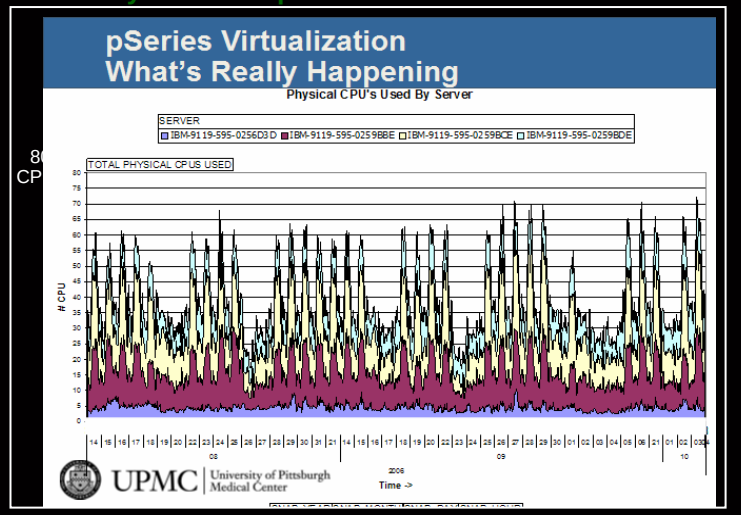
UPMC
University of Pittsburgh Medical Center

- Production:
  - 4 p595 Servers – 64 CPUs 512 to 640 GB
  - All 256 CPU allocated
  - 222 LPARS

- Disaster Recovery Site
  - 5 p570 Servers – 16 CPUs 64 GB Memory
  - 31 LPARS

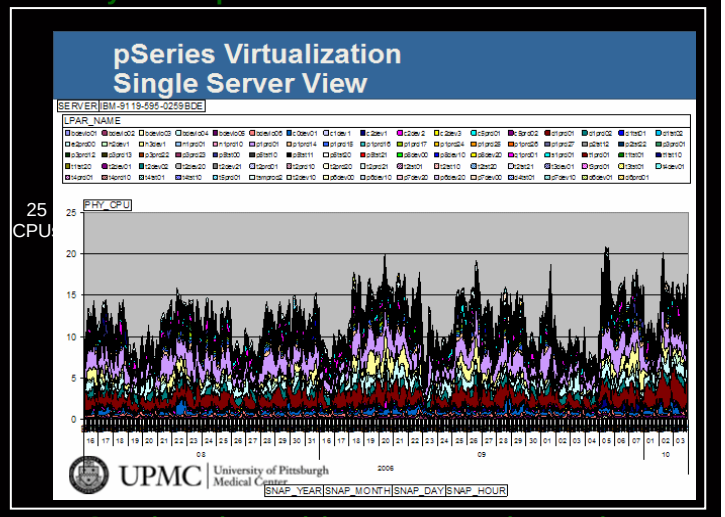
- Implemented full virtualization:
  - ☑ Reduced Peak requirements by over 50%
  - ☑ Provides over twice the performance capability
  - ☑ Provision new virtual servers in hours
  - ☑ Created 184 spare virtual CPU’s for new workloads

Summary of all four p595’s



With full virtualization peak usage dropped to 72 cores across 4 systems

Summary of one p595



84 virtual machines now only peak at 21 cores on a 64 core system

IBM’s EnergyScale 技术—全面的解决方案

Virtualization

- Increase utilization rates
- Reduce number of server, storage, network devices
- Create shared infrastructure

Active Energy Manager

- **Helps companies meter, control, even cap their power usage**
- **“Cruise control” for power consumption of servers**
- **Available on all IBM servers**

Partition Mobility with P6

- Migrate workloads to eliminate hot spots
- Move work off underutilized systems to conserve power

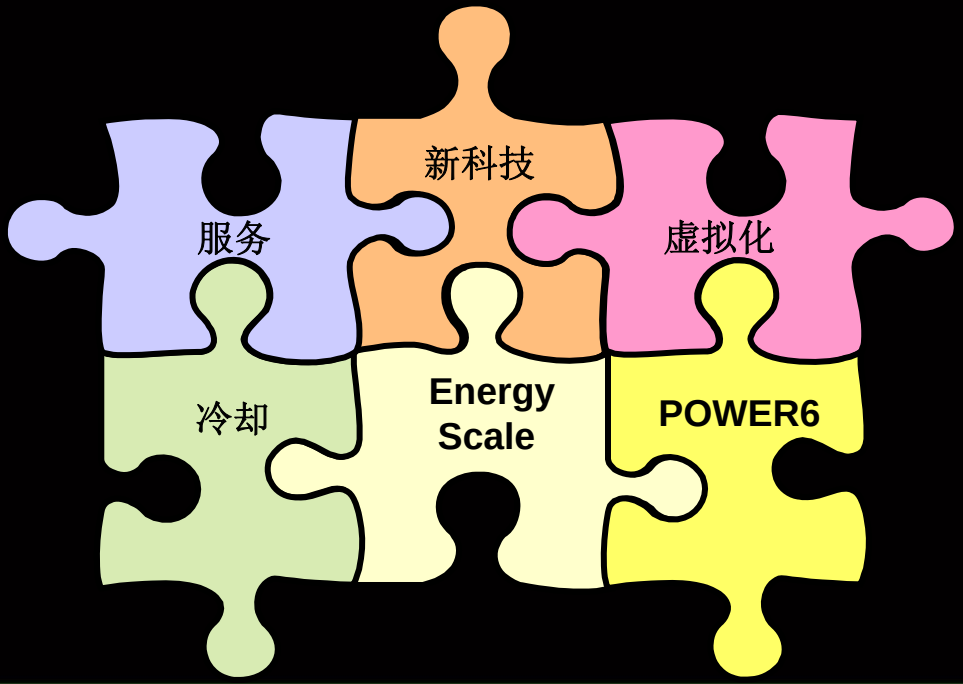
IBM Virtualization Solutions

PowerVM™

IBM的绿色计算....

- 虚拟化技术: 提高利用率: 更少的服务器, 更少的处理核
- 服务: 高效率的数据中心
- 新科技: 性能/ 瓦特.
- POWER6: Nap 模式 / POWER 内存 / etc.
- 冷却: 背门热交换器
- EnergyScale: 监测能耗 / 按虚减少能耗

IBM is the only vendor who can put all of the pieces together



IBM 是帮助你实现节能减排的伙伴

- **IBM** 提供前期顾问服务、领先的硬/软件产品、方案实施服务等可以帮您跨越目前数据中心建设的限制。
- **IBM** 与客户一起努力，无论大客户还是小客户，使绿色数据中心真正为客户产生业务效益。
- **IBM** 是公认的节能的市场领导者。

POWER6