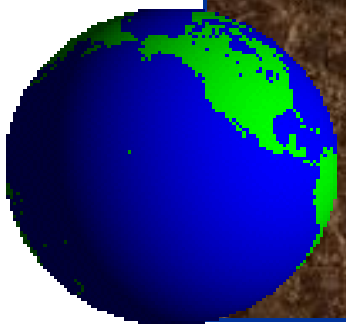
中国科协学术论坛

INSTITUTE OF COMPUTING TECHNOLOGY

节能优先 —
芯片与信息系统设计思想的大转变



中科院计算所
李国杰
2011年8月10日

节能需要考虑一些全局问题

因节电改变了信息产业的格局
--从RAM公司的崛起谈起

- 去年全球有61亿颗芯片采用ARM架构（ARM inside），芯片市场占有率为28%，全球每四颗芯片，就有一颗来自ARM血统。到2020年ARM累计芯片销售量将达1500亿颗。微软宣布支持ARM，垄断30多年的Wintel联盟瓦解

- ARM成功的秘诀，就是追求**“省电”**
- **节能的需求**促使**信息产业**产生**颠覆性的变革**。

- 苹果、剑桥、牛顿、ARM; 去年世界大学排行榜上，剑桥首度击败哈佛，登上榜首。

节能的出路在提高信息设备能效

$$E = \sum(P_i \cdot T_i / U_i); \quad i \text{ from } 1 \text{ to } M$$

- E：社会所有信息设备的总耗电量
- M：社会信息设备的总拥有量 ↑
- Pi：单个信息设备的性能 ↑
- Ti: 单个信息设备的使用时间
- Ui: 单个信息设备的能效 ↑ ↑

中国计算机与服务器的保有量

计算机能广泛应用于设计优化,制造优化,配送优化以及各种产品运营优化等方方面面

2007-2015年中国计算机保有量及预测

2007-2015年中国服务器保有量及预测

统计分析显示,计算机的应用直接或间接地提高了能源效率。在过去的1949到2006年间,美国信息技术所消耗的每千瓦/时电能,在整体经济范围内约能节省8.6千瓦/时电能。

本页及以下3页内容引自CCID Consulting报告，供参考

中国计算机与服务器的总功耗

计算机及服务器的使用,在提高各行业生产效率的同时,也成为电能消耗的主要设备。

- 以每台计算机功耗300w、每天工作8小时计算;
- 2009年,中国计算机产品上消耗的电能约为1927亿千万时;
- 这一电能消耗是当年全国总发电量的5.3%,相当于2.5座三峡水力发电站的年发电量。
- 在现有技术情况下,随着保有量的不断增长,到2015年,中国计算机产品上消耗的电能将高达5500亿千万时;

- 以每台服务器功耗800w、每天工作24小时计算;
- 2009年,中国服务器产品上消耗的电能约为261亿千万时;
- 以服务器应用为基础的数据中心(IDC)上的能量消耗,约是其服务器上消耗电能的3倍,初步估计,2009年,全国IDC中心消耗的能量约为783亿千瓦时;
- 在现有技术情况下,随着保有量的不断增长,到2015年,中国服务器产品上消耗的电能将高达656亿千万时,IDC上消耗的电能则超过1900亿千瓦时;

计算机领域未来5年可能省多少电？

随着半导体技术的不断创新及其在计算机和服务中的深入应用,未来,计算机及服务器的电能效率将显著提升,电能消耗量则随之大幅降低。与此同时,计算机及服务器性能的提升以及其在各行各业中的进一步普及,也将通过提高单位时间生产效率的方式,为全社会能源效率的提升作出重要贡献。

产品	2009年		2015年	
计算机	保有量	2.2亿台	保有量	6.3亿台
	电能消耗	1927亿千瓦时	采用先进半导体技术后的电能消耗	2750亿千瓦时
	-		现有技术情况下的电能消耗	5500亿千瓦时
			能源效率提高	100%
服务器	保有量	373万台	保有量	936万台
	电能消耗	261亿千瓦时	采用先进半导体技术后的电能消耗	406亿千瓦时
	-		现有技术情况下的电能消耗	656亿千瓦时
			能源效率提高	62%

注:采用先进半导体技术后的电能消耗指的是,采用高阶工艺制程、多核处理器芯片、智能电源管理芯片等一系列节能技术后,2015年,计算机或服务器上所消耗的电能;
现有技术情况下的电能消耗指的是,假设到2015年,计算机及服务器仍采用2009年时间点上的半导体技术,计算机或服务器上所消耗的电能;

通信领域未来5年可以省多少电？

高速发展的半导体技术是深入推进通信行业绿色节能的基础

更加先进的半导体器件与配套软件的开发与应用，可使基站覆盖范围扩大20%-40%，同时大幅减小主设备体积与重量，主设备功耗下降40%-60%。

当主设备功耗低于730W，电力供应上可直接利用太阳能、风能、沼气等绿色能源，节省投资成本，彻底消除基站和配套设施带来的环境污染。

当功耗降低到550W以下，则无需空调辅助降温。若再用自然散热，可降低配套设备绝大部分能耗。

产品	2009年		2015年	
移动通信基站设备	保有量	103.7万台	保有量	185万台
	电能消耗	272.5亿千瓦时	采用先进半导体技术后的电能消耗	292亿千瓦时
	-		现有技术情况下的电能消耗	486亿千瓦时
			能源效率提高	67%

在网络上每传送1bit的功耗每年降低30%左右

信息为什么这么“重”？
—解决网络功耗问题的联想

- 假设要传送200TB的数据从北京到西安，按2TB的硬盘一公斤计算，大约 100KG重, 从邮局寄包裹或交铁路物流托运，运费不会超过**500元**。
- 若是用租用1Gps的专线，年租费70－180万元，按天算2000－5000元，每天满打满算可传送约10TB，需要20天，租金需要2－10万元。
- 这相当于火车运送硬盘的“原子”只有100公斤重，但送过去的“BIT”按运费算超过10吨，BIT比原子“重”100倍。
- 尽管光纤的带宽不断增加，但目前全世界带宽最宽的网络是邮政或物流系统。从光纤上传送数据的功耗与要传送的数据量及光纤带宽究竟是什么关系，需要深入研究

从节能的角度权衡通信和存储

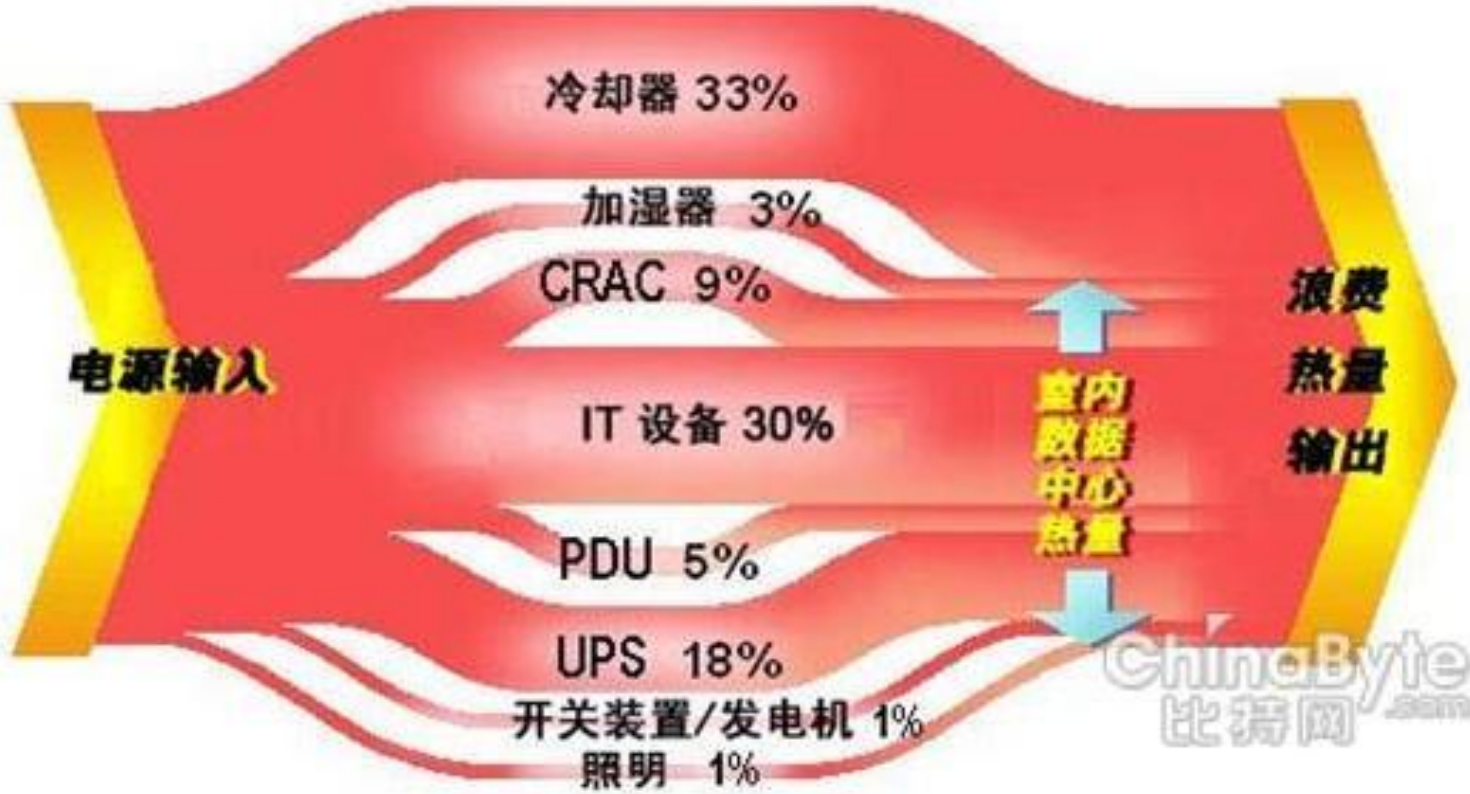
- 共享存储和数据通信都能起到信息交流的作用，有些情况下存储可以替代通信，CDN的大缓存就能起到减少远程通信的作用。**存储和通信都需要能量，需要做Tradeoff**。
- 由于个人存储有大量冗余，总体上讲，云计算比个人分散存储节省存储空间，也可能节省能量。但云计算明显增加了通信量，从全社会用于信息处理、存储和通信的能量最少的目标出发，**需要从总体上做全局优化分析和设计**。
- 尽可能**减少冗余的存储和通信**是节能的重要途径。
- **增加Cache容量**比增加用于逻辑部件的晶体管数更**有利于节能**，因此近年来更多的芯片面积将用于Cache。

节省能量与节省带宽不能两全

- 我们不仅需要预测通过一个网络能够传输的最大信息量，而且还要预测传输这些信息量需要消耗的最小能量。贝尔实验室的研究表明，当前的ICT网络具有将能效提升10000倍的潜力。
- 贝尔实验室的研究得出结论：**能量效率和频谱效率**之间有一个折中的关系，即较低的能量效率（焦耳/比特）将产生较低的频谱效率（bps/Hz）。即对同样的数据传输率，**较低的能量消耗就需要更多的频谱带宽**。
- 对固定数据传输率而言，能量的节省将转变为对带宽的需求；降低蜂窝小区尺寸是目前减少每比特发射能量最有效的办法，需要的代价是通过增加天线数量增加带宽。用1个天线时，平均每比特发射能量是7nJ；而用4个天线时，就小于1nJ。

提高数据中心的能量效率（PUE）

图 1– 典型数据中心的电能流向



数据中心散热的新招

尽早谋划如何解决信息化的节能

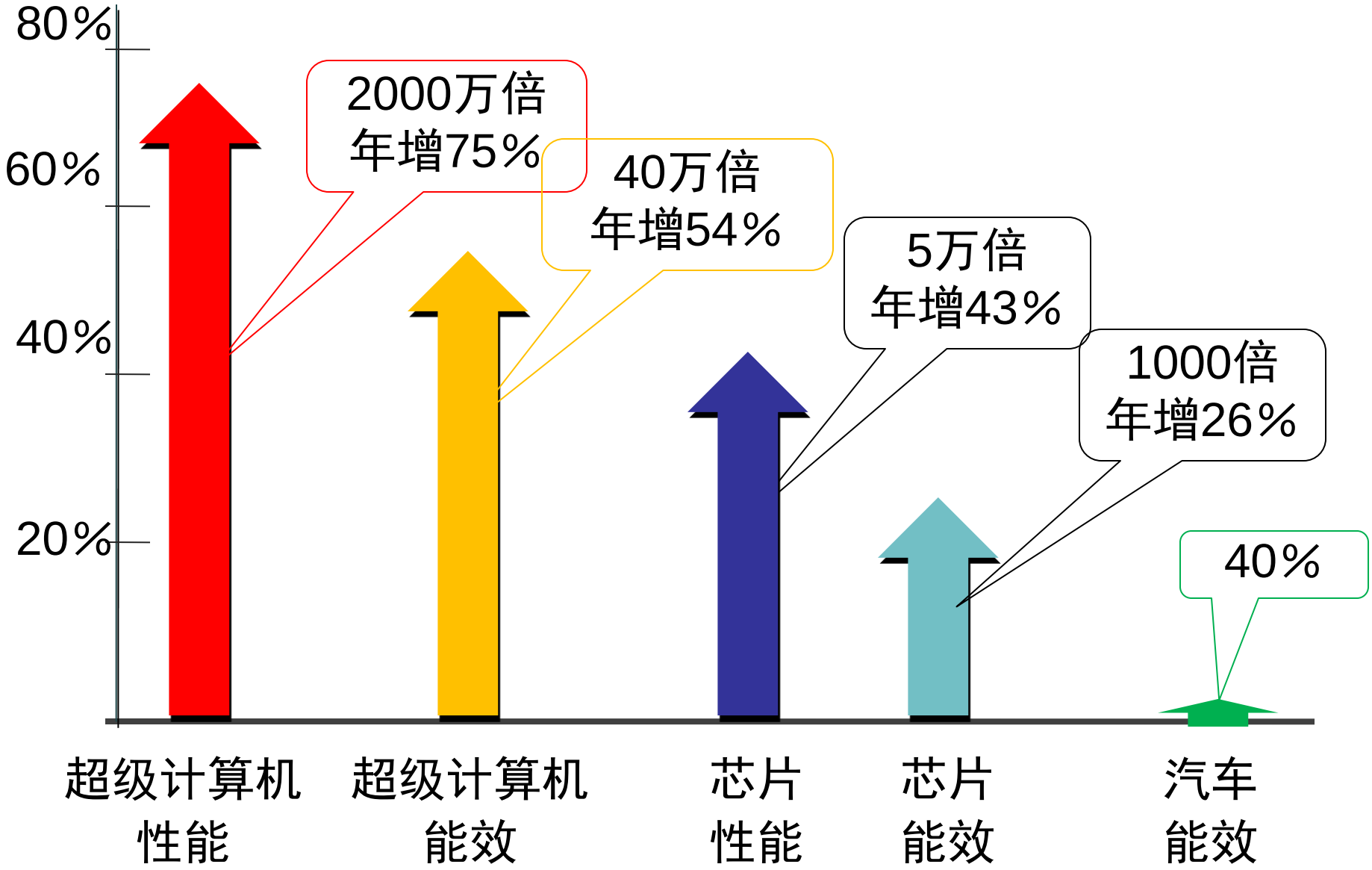
- 专家估计，未来5年数据存储需求的复合年增长率为35%至65%。Gartner 认为，2009－2013的5年内，用户购买的存储能力将是2008年购买量的**20**倍。
- 按“数字化信息总量与全球平均水平相当”这个假设来估计，2010年我国的数字化信息总量约为**205.2EB**。按（35%-65%）的中间值48.5%预测，估计我国数字化信息总量的增速，则2020年中国数字化信息的总量可能达到**10ZB**(10^{22}B)，约100亿TB，人均接近**10TB**。
- 信息化对能量需求的可怕之处是其增长速度惊人，如果不早做谋划，任其盲目发展，东南沿海和大城市的数据中心 热必将加剧我国的电力紧张。

节能的根本出路在基础研究

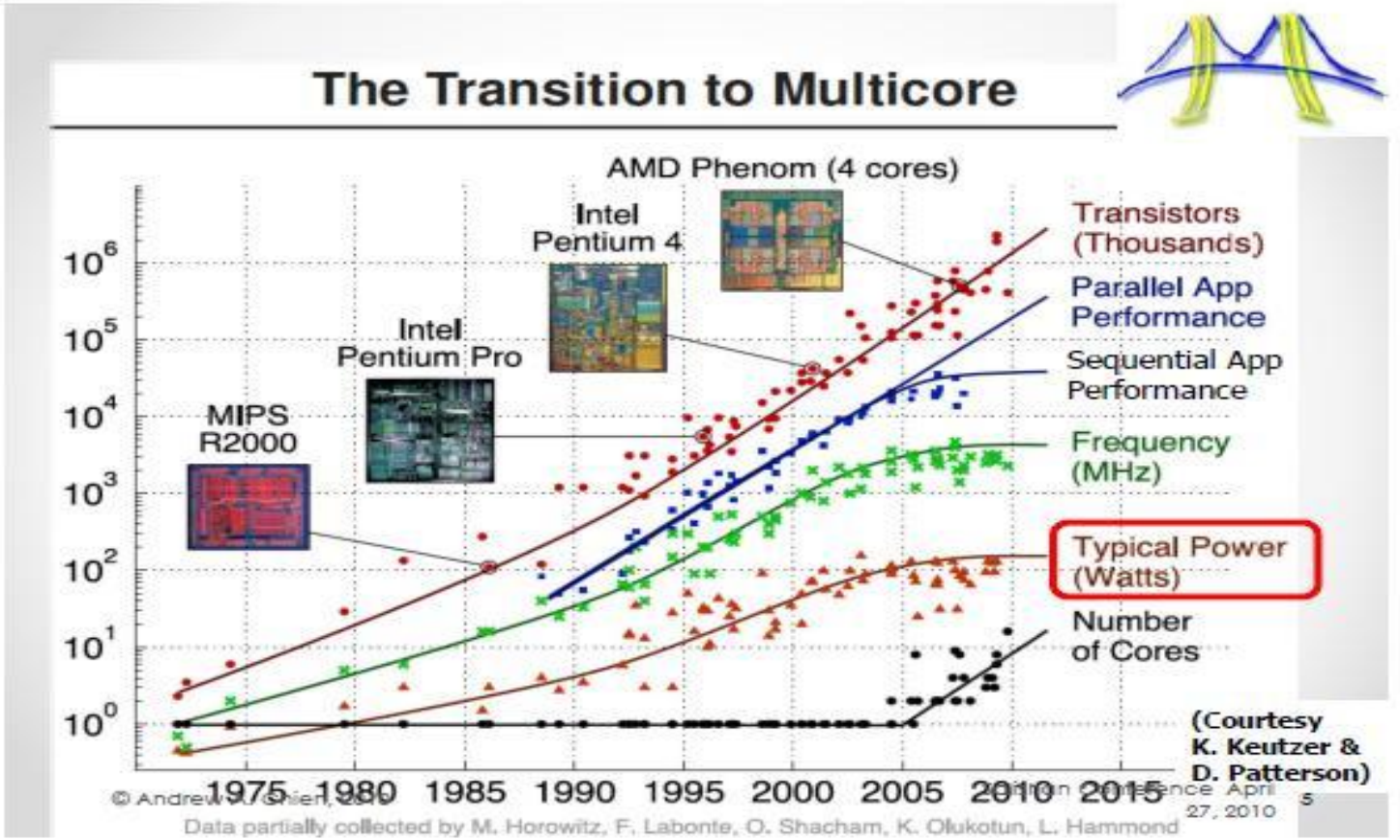
提高功耗效率已成为首要目标

- 集成电路、信息系统的性能和使用规模都在指数性增长，导致人类使用的信息系统的总能量急剧增长。**信息系统的能量效率（单位功耗获得的性能）的提高必须高于系统性能增长的速度**，才能实现节能的目标。因此，提高功耗效率已变成比提高性能更优先的设计目标。
- 目前提高信息系统功耗效率的途径很多，从芯片到系统设计，从操作系统到应用软件都可以为节能做贡献，需要全局优化。
- 从另一个角度看，途径多也说明还没有找到信息系统节能的主要突破口，还需要从原理上深下功夫。

计算机性能和能效
30年增长倍数和年增长率



单芯片的功耗已不能再增长
今后提高芯片性能只能靠提高能效



降低系统功耗的多种途径

如何降低运算所消耗的功耗？ $P = KC_{out}V_{dd}^2f_{clk}/2 + I_{sc}V_{dd} + I_{leak}V_{dd}$

Layer	Technique	Reductions	Factors	References
System	System shutdown	41%~99%	V , k	[CSB94,CIC94]
	Dynamic voltage & frequency scaling	10%~73%	V, f	[SEO05]
	Algorithm selection	33%	k	[OY94]
	Compiler optimization	13%~20%	k	[Lee00]
Architecture	Data representation	13%~32%	k	[yu02][STD94]
	Parallel processing with low voltage	51%~80%	V, C, f	[CSB92]
	Cache design	20~80%	V, C, f	[Yang02] [BAF94,PR95]
	Bus encoding	15%~48%	k	[Lyuh02]
	Operand isolation	30%~40%	k	[Banerjee06][Munch00]
Logic	Logic synthesis	<70%	C, k	[Hsu02][IP94][TMA95]
	Clock gating	20%~75%	k	[Li02][Monica03]
	Technology mapping	<47%	C, k	[LM93][TAM93]
	Path balancing	9%~41%	C, k	[Kim01][BCH94]
Circuit	Low swing clock	30%~63%	V	[ELE00][HK98]
	MTCMOS,VTCMOS,DTCMOS	20%~80%	I_{leak}	[Li99][Far97][Tad96]
	Power gating	<100%	V, I_{leak}	[David02]
	Device stacking	1%~56%	I_{leak}	[Halter97][Rahul03]

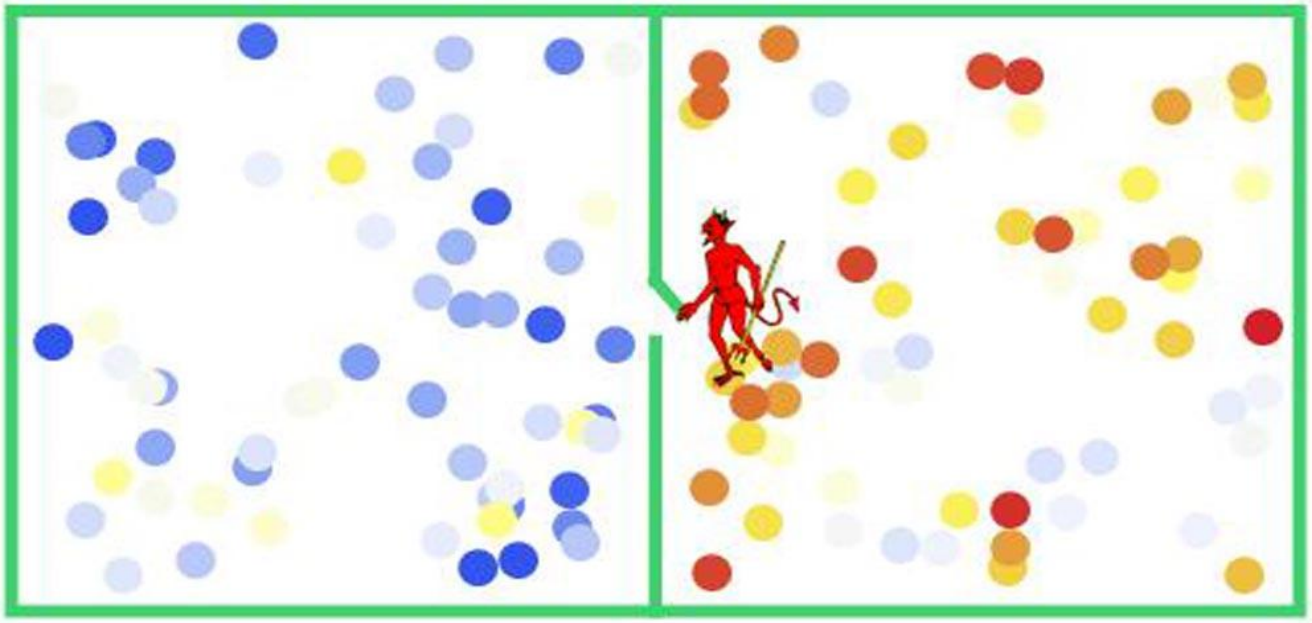
需要研究一些最基础的问题

- 节能问题不只是一个渐进改进和优化的技术问题，本质上是一个涉及能量和信息转换的基础科学问题。
- 苏联物理学家朗道推导出：一位0-1变换需要的能量是**kTln2（约10^{-21}焦耳）**，这是热力学第二定律所允许的最小量，称为朗道极限。从理论上讲，熵减少必须消耗能量，可逆计算机（熵不变）原则上不耗能。最新研究显示，清除信息无需任何能耗理论上可以实现，信息清除的成本可由另外一种存储体系如自旋角动量来支付。
- 人脑是最节能的信息处理与存储系统，从根本上解决信息系统的节能问题需要从脑科学得到启示。光计算机、量子（超导）计算机、分子（DNA）计算机等非传统计算机的能效都比电子信息系高几个数量级，但估计20年内还取代不了电子系统。

终极能效计算机

- 2006年时，法国巴黎圣母大学的研究人员第一次用磁性纳米粒子成功地演示了一次逻辑操作，发现这种电路能在朗道极限能量下运行。
- 最近，加州大学伯克利分校一位研究生布赖恩•拉姆森用一种宽约100纳米、长200纳米的磁铁制作了磁性存储和逻辑设备，当多个纳米磁铁结合在一起时，通过两极间的力相互作用，能够实现简单的逻辑运算。这种芯片每次操作仅耗用**18毫电子伏特（约10^{-21}焦耳）**能量，比目前计算机每次操作的耗能低**100万倍**。
- 2010年能效科学研究中心从美国国家科学基金会获得了**2500万美元**拨款，其中一个目标就是建造在朗道极限能量下运行的计算机。向朗道能效极限进军无疑还要克服很多困难，用于产生磁场、擦除或翻转纳米磁铁的极性的电流也会消耗许多能量。

信息直接转化为能量的实验



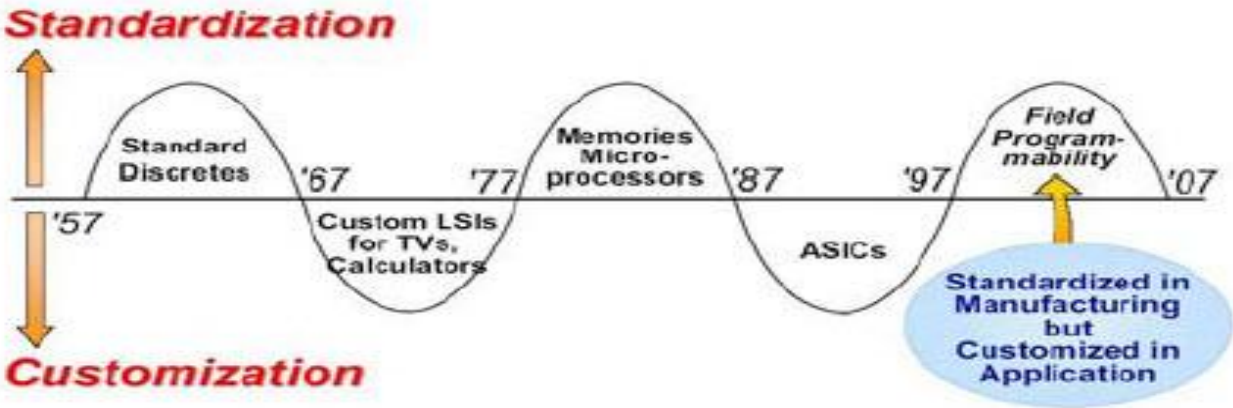
日本研究人员2010年11月发表报告称，他们在实验室中让一个纳米小球沿电场制造的“阶梯”向上爬动，爬动所需的**能量**由该粒子在任何给定时间朝哪个方向运动这一**信息转化**而来，这意味着科学家首次在实验室实现了**信息到能量的转化**，验证了约150年前英国物理学家詹姆斯·克拉克·麦克斯韦提出的“麦克斯韦妖”这一设想。

从节能角度看通用和专用

从节能出发改变芯片结构设计

- 在信息系统和芯片的发展史上，通用和专用系统（芯片）总是呈现周期性的交替发展趋势。
- 由于针对某些领域应用的专用系统（芯片）的功耗效率较高，集成多个专用部件的异构芯片将成为今后芯片发展的主要方向。
- **90/10** 优化原则已不再适用，为了节省能量，更好的途径是设计多个加速器，每个加速器只针对10%的情况优化,这种方法称为 **“10×10 优化”**。 每一时刻只有10%的晶体管活跃，其他90%晶体管不活跃。

半导体产业的牧村周期



Source:Electronics Weekly, Jan.1991

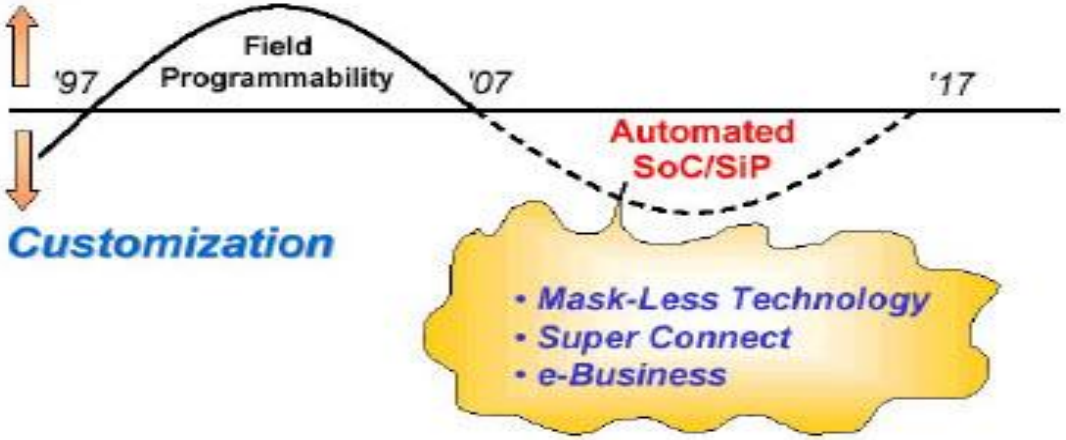
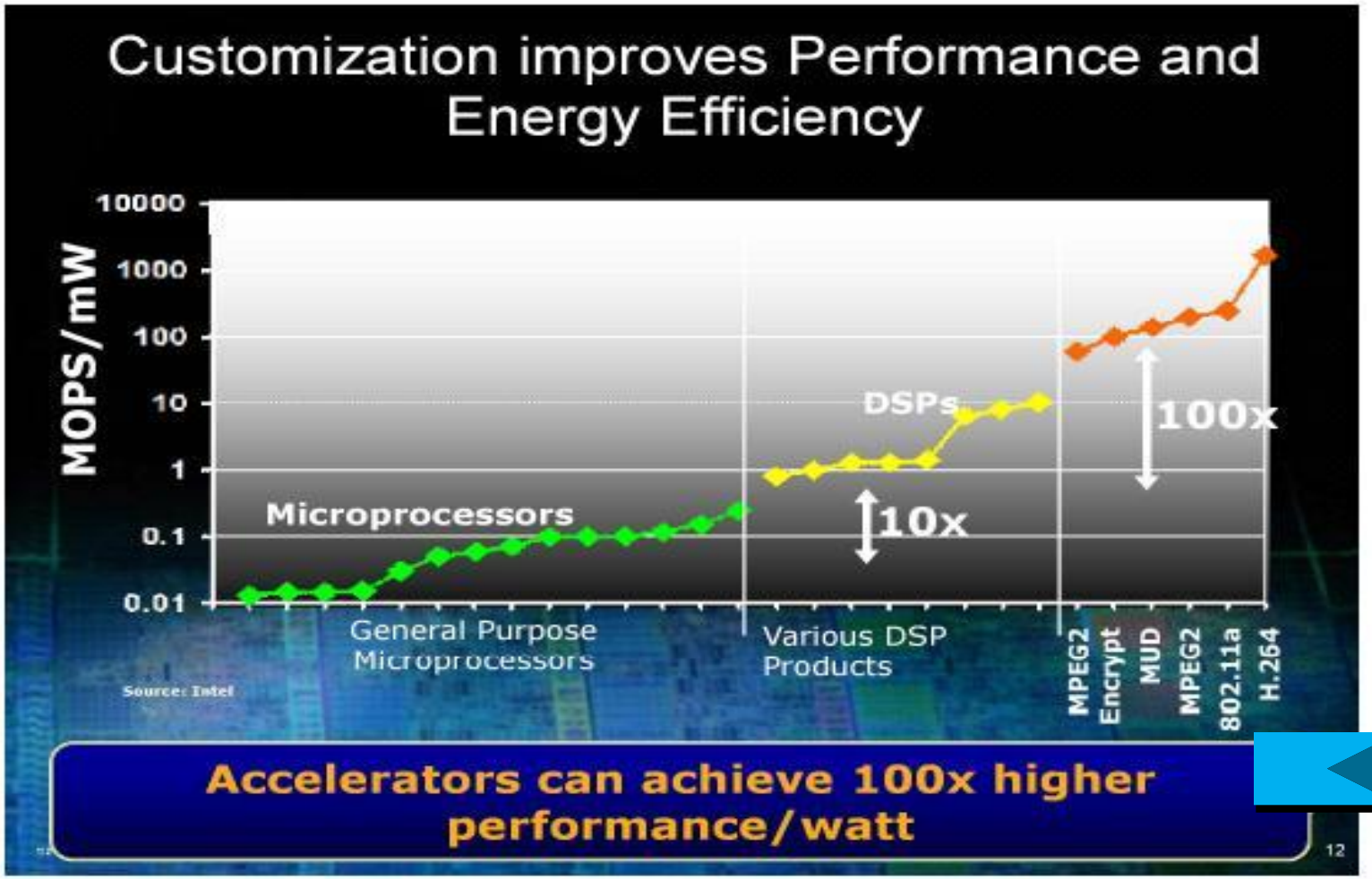


图 1 牧村浪潮及其首次修订

加速部件可以提高能效100倍



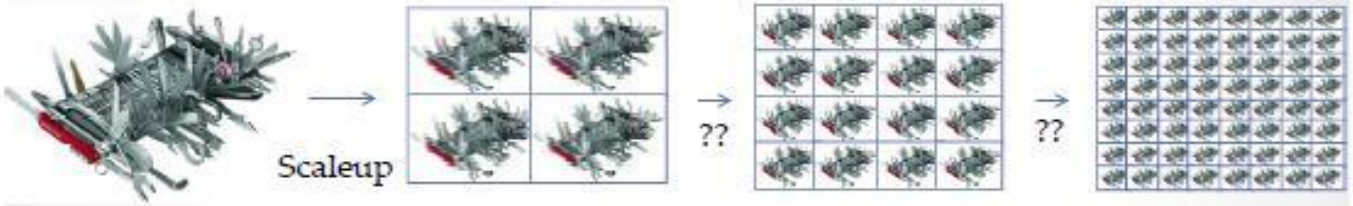
90/10优化模式

90/10 Aggregation leads to Energy Inefficiency

- Microprocessor complexity grows with each new optimization/feature
  - Instruction set compatibility (binaries must run)
  - Performance monotonicity (apps must not slow down)
  - New functionality – double fp, mmedia, virtualization, crypto, etc.

目前主流的处理器系统结构(Multi-core）
——集成几个复杂的CPU核，功耗高

Approach #1: Integrate Lots of Traditional Cores

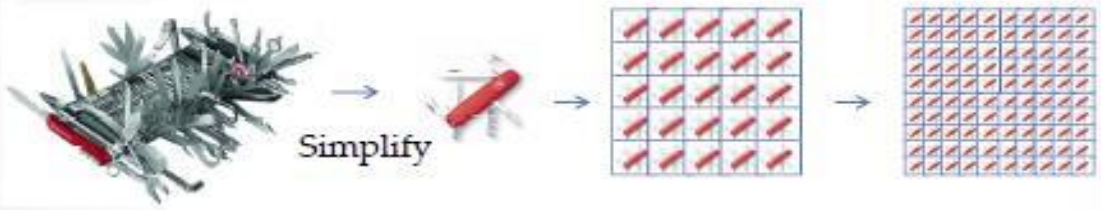


- Mainstream processors market approach (preserve software compatibility and performance monotonicity)
- IBM Power, Intel Xeon, AMD Opteron are delivering complex, high performance cores into the mainstream market and scaling the number of cores

Won't get too far, low energy efficiency limits scaling

节能的Many-core：集成大量简单的CPU核

Approach #2: Simplify for Energy Efficiency, then Scaleup



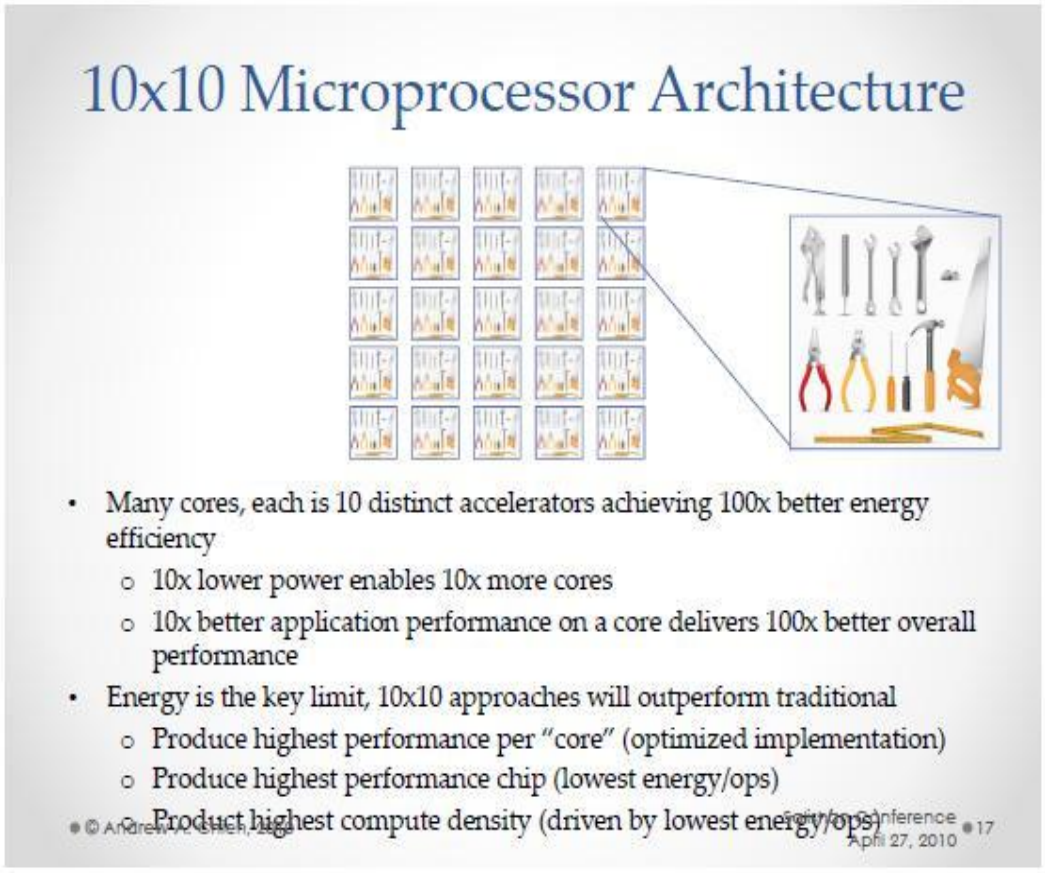
- Simplify: strip out micro-architecture techniques; adding execution restrictions (SIMD, hierarchical control, restricted data access)
- Examples: Blue Gene/L, Blue Gene/P, Sequoia
  - And GPUs…
- Caution: Interconnect energy scaling

一种新的高能效芯片设计思路

An Alternate Vision for Energy Efficient Computing

- Create nodes that are a collection of customized, special tools which are designed to purpose
  - A set of tools, each high performance and energy efficient for its computational structure
  - Tools complement each other in purpose and use (different specialties)
- Only one tool used at a time! (others powered off)
  - That tool achieves high performance and energy efficiency…

10X10多处理器系统结构



- 集成10个不同用途的加速部件，提高能效100倍
- 10倍低功耗可增加10倍IP核
- 由于每个核专用，性能可增加10倍。

未来三种芯片结构谁是主流？

What is the future of computing?

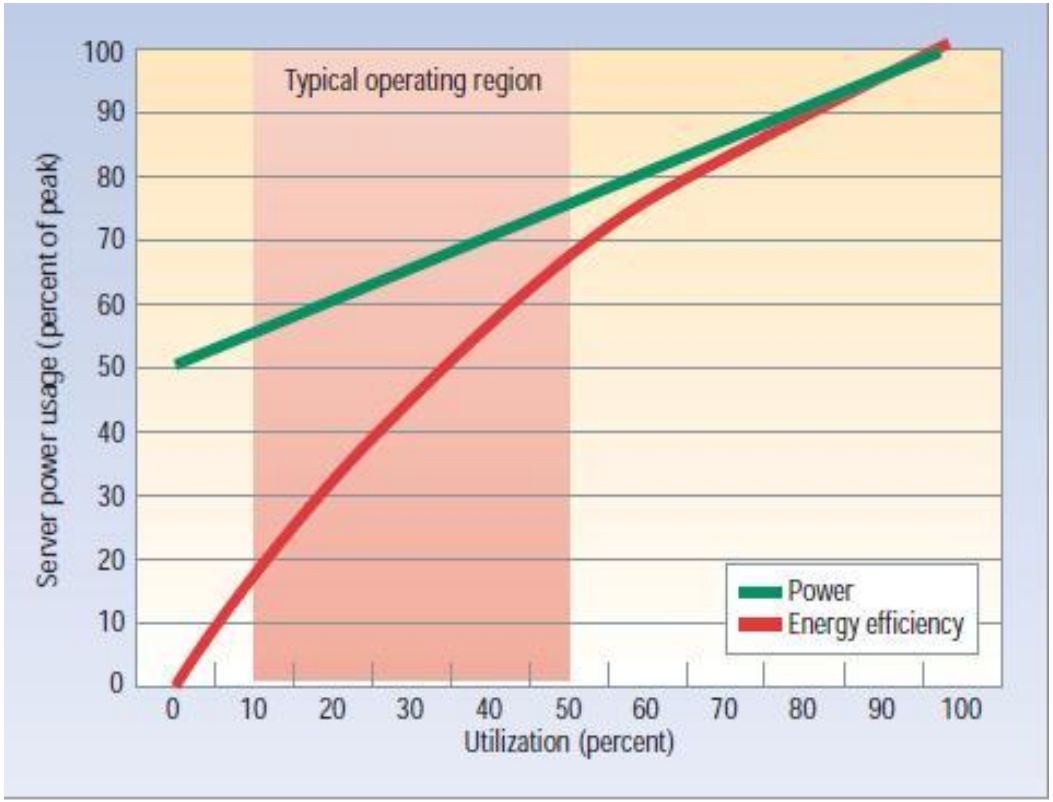
OR

OR

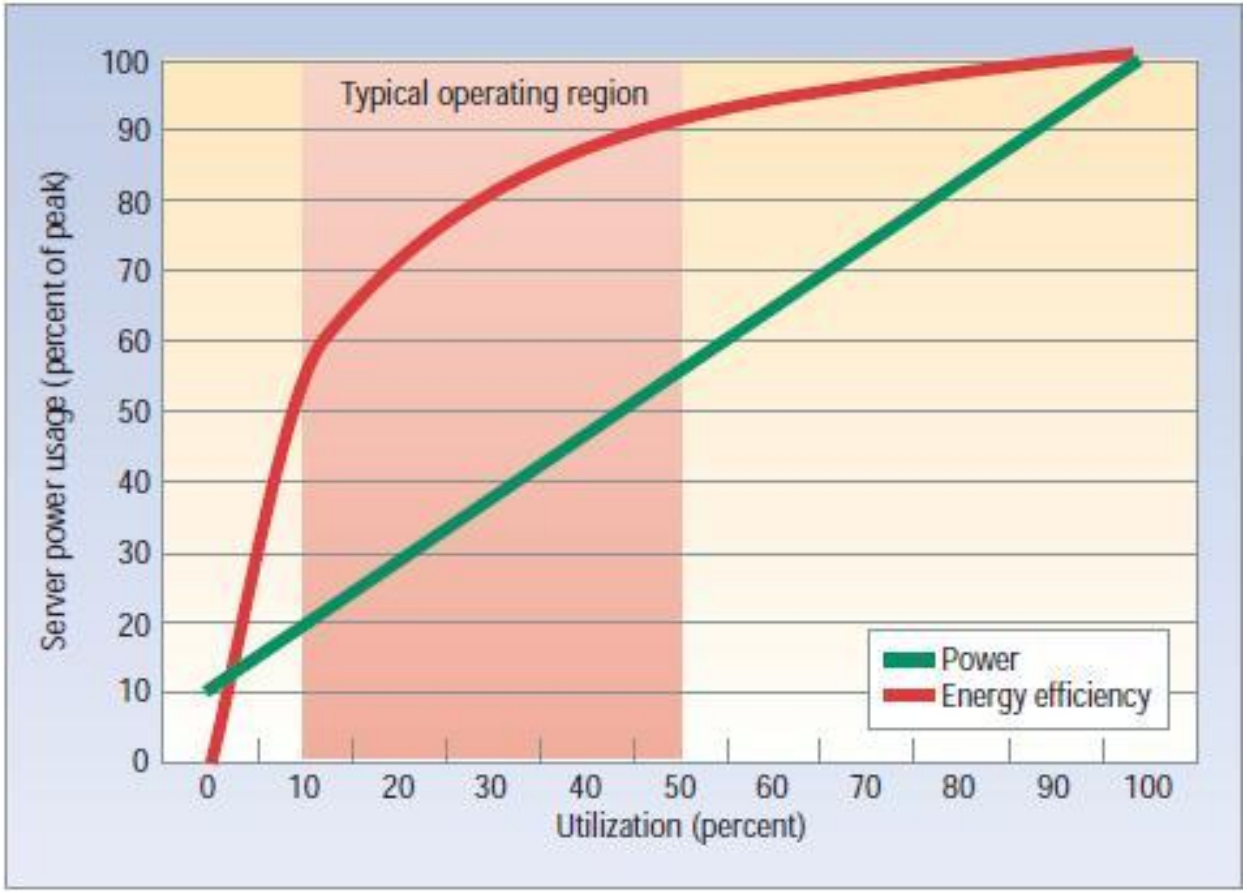
降低空载功耗和
设计功耗匀增系统

功耗匀增（energy-proportional）问题

- 多数信息系统运行时的**实际功耗往往并不正比于负荷和性能。**
- 服务器、磁盘、网络交换器在**空载**时分别要消耗其峰值功耗的**30%、75%和85%**
- 如何设计**energy-proportional** 信息系统是需要高度重视的问题。

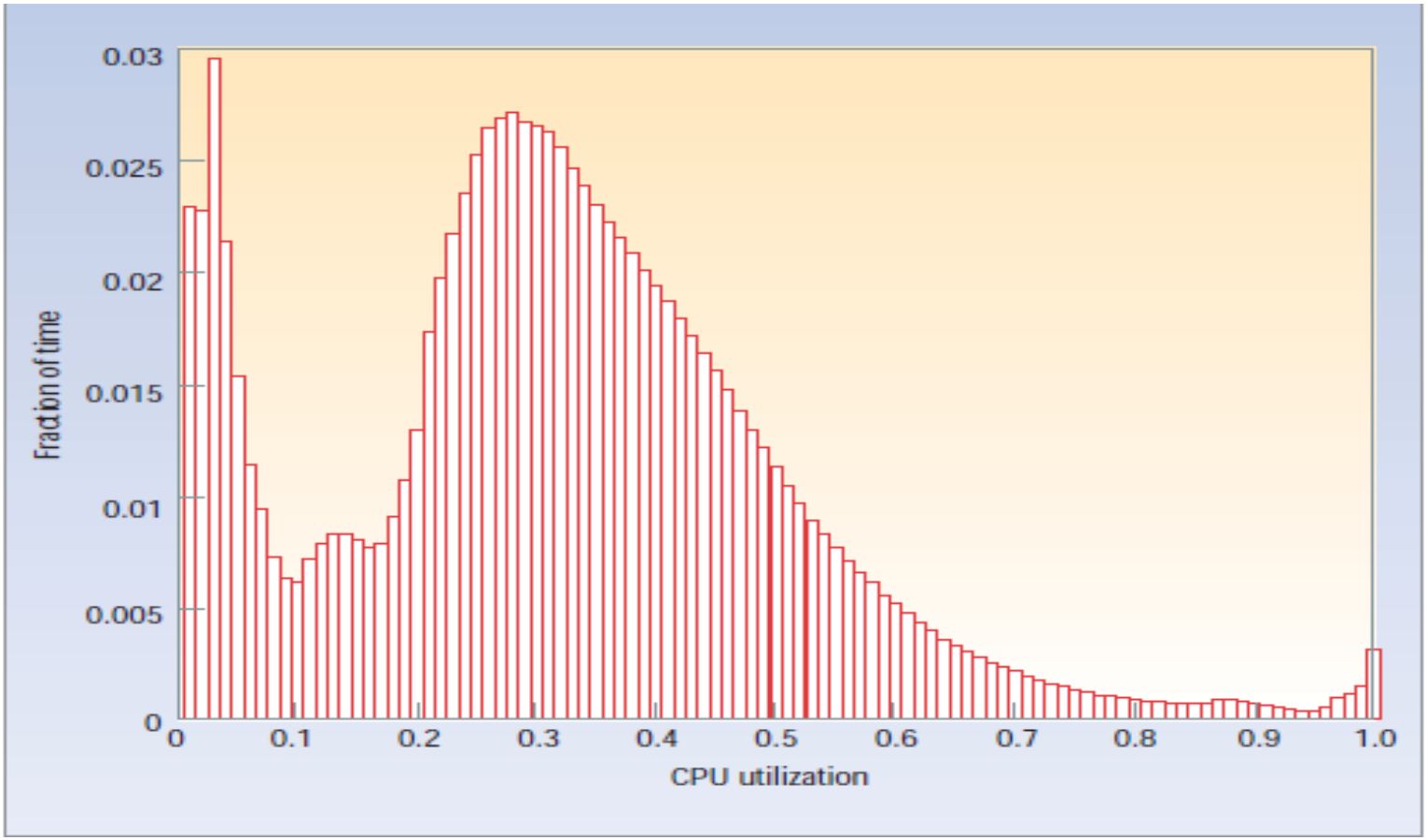


节能的服务器能效变化模式



- This server has a power efficiency of more than 80 percent of its peak value for utilizations of 30 percent and above, with efficiency remaining above 50 percent for utilization levels as low as 10 percent.

服务器的工作模式不同于移动设备



- 服务器不同于手机，很少完全休息或满负荷工作，大部分时间服务器的利用率在10%至50%之间。

Energy-Proportional Computing

- 当前的服务器的最低能效区域对应对常用的操作模式，解决这一问题需要重新思考部件和系统的设计。
- 能效必须与计算机性能增加同步，才能避免不断扩大机房面积。
- 服务器不可能利用现在的节能模式达到移动设备一样的高能效。我们必须研发一种机器，它的**能量消费正比于负载和性能**。理想的能量匀增机器在空载是应该功耗为零，在极低负载是几乎没有功耗，随着负载增加功耗线性增长。
- ——内容摘自 **Luiz André Barroso and Urs Hölzle（Google公司），“The Case for Energy-Proportional Computing”，2007年12月 IEEE Computer**

实现功耗匀增系统的途径

- 功耗匀增系统可以大量节能，提高能效一倍以上。 实现功耗量匀增系统需要明显改进内存和硬盘的能量使用概况（energy usage profile ）
- 采用**大规模并行、异构的IP核和加速部件**是提高能效的部件途径
- 为了实现功耗匀增的计算，通过**软件和硬件的密切磨合**达到有效的数据配合（data orchestration）十分关键，.
- 系统结构将从同构走向异构，慷慨地使用晶体管去配合面向客户应用的硬件。需要**增加软件的并行性**来适应异构加速部件和面向客户的应用。

提高动态功率范围

- 人的平均能量消费是120W，休息是大约70W，但是人剧烈运动时可达到1KW(保持10分钟左右），运动员甚至可达到2KW。
- 移动和嵌入式CPU空载功耗不到峰值功耗的10%，台式机和服务器处理器在很低负载时消耗峰值功耗的30%左右，动态范围70%。其他部件的动态功率范围更小：DRAM 50%，磁盘驱动器25%，网络开关的动态功率范围只有15%。
- 磁盘等设备从从休眠模式到活动模式的转换有延迟和能量开销，这种开销可能显著降低系统的性能，因此休眠模式控制在**毫秒**级别。
- 提高动态功率范围是解决节能问题的主要方向之一。

请批评指正！