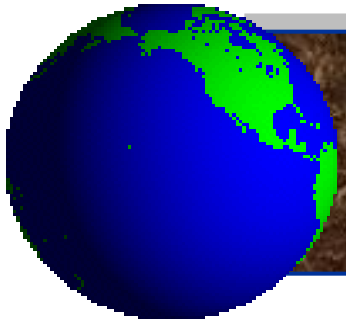
INSTITUTE OF COMPUTING TECHNOLOGY

加强计算机系统研究的一些考虑



李国杰，中科院计算所

2010.12.14

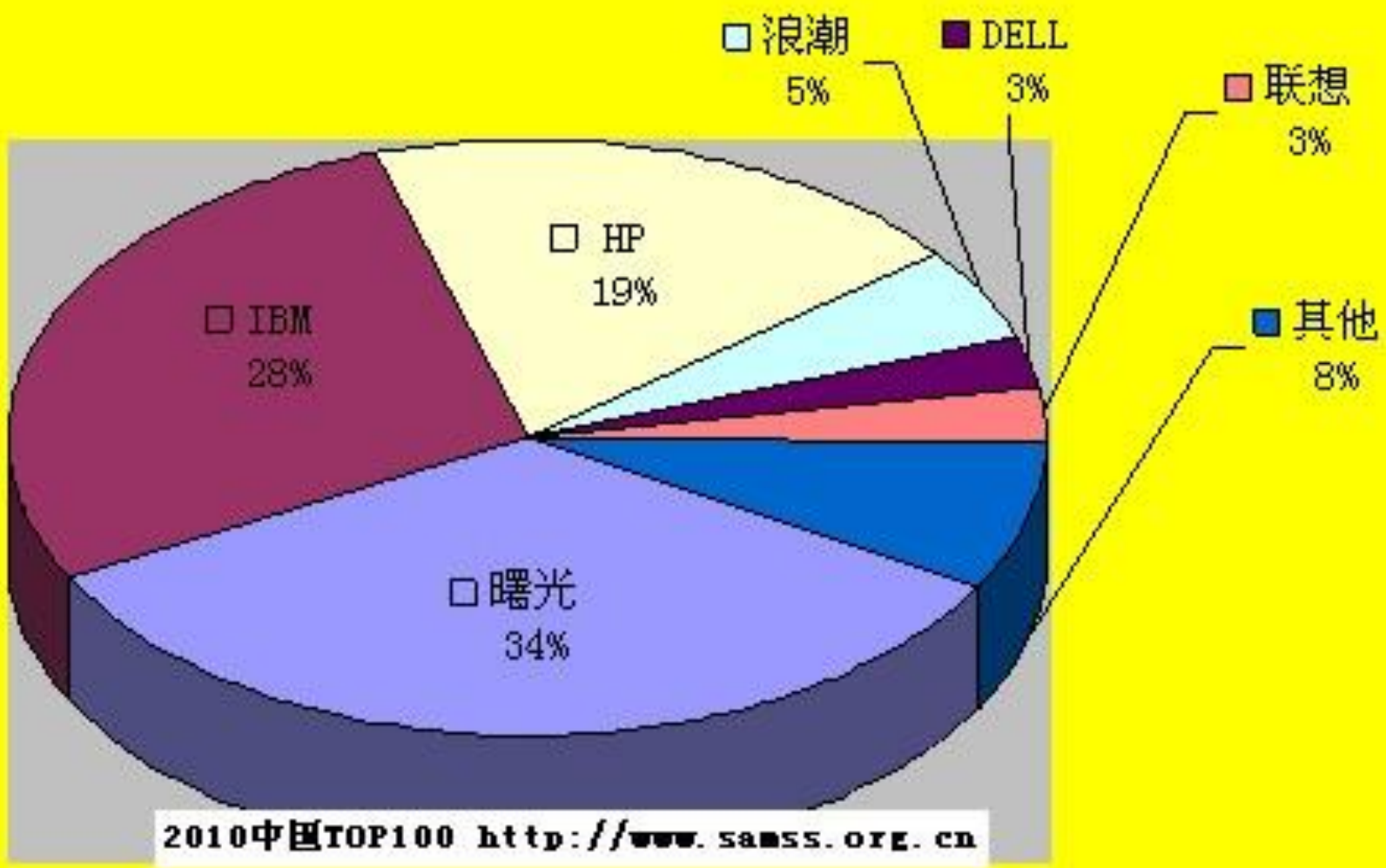
对我国计算机系统研究水平的判断

目前排名世界第一的天河－1A

曙光星云超级计算机

在一次HPC竞赛项目中
升起两面中国国旗

曙光超级计算机在国内TOP100 HPC中占34台，超过IBM和HP



IBM明年将推出10Pflops 的 Bluewater

计算机领域国家重点二级学科

081201计算机系统结构

- 华中科技大学

081202计算机软件与理论

- 吉林大学
- 复旦大学
- 中国科学技术大学
- 武汉大学

081203计算机应用技术

- 东北大学
- 东南大学
- 浙江大学
- 安徽大学
- 四川大学
- 西北工业大学

- 计算机系统结构方向还没有设立国家重点实验室

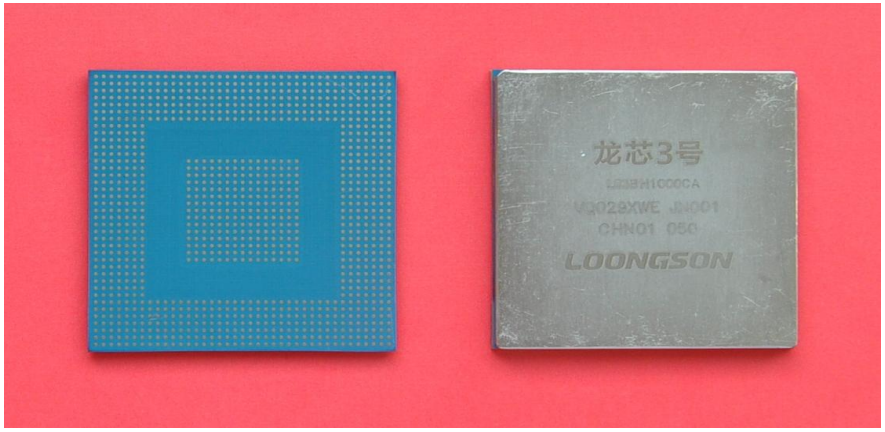
2006-2010年我国在计算机体系结构领域国际顶级会议和期刊上发表论文情况

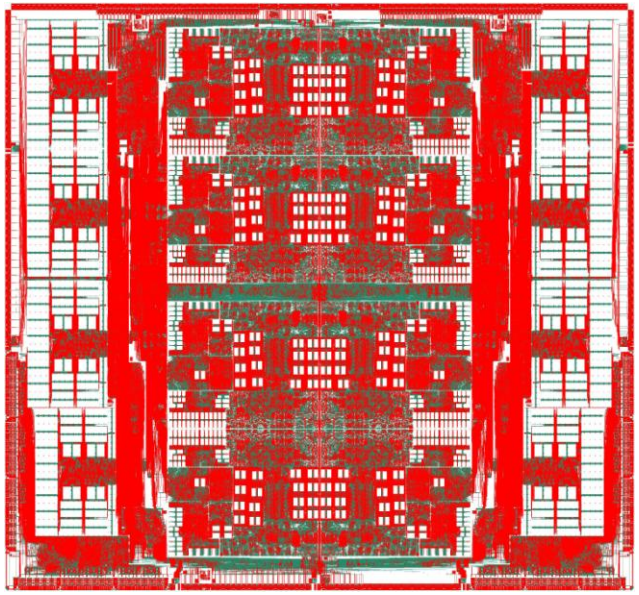
顶级会议/期刊	大陆发表论文数	中科院计算所论文数	国内论文占比（总数）
国际计算机体系结构会议 (International Symposium on Computer Architecture (**ISCA**))	4篇	2篇	2%（198）
国际微体系结构会议 (International Symposium on Microarchitecture (**MICRO**))	2篇	1篇	1%（214）
国际高性能计算机体系结构会议 (International Symposium on High-Performance Computer Architecture (**HPCA**))	3篇	2篇	2%（159）
IEEE 微体系结构期刊 (IEEE Micro) IEEE CS硬件领域排名最高期刊 影响因子3.205	3篇	2篇	～1%

- 三个顶级会议长文录取率约为**18%**

龙芯3B达到每秒1280亿次浮点运算

- 8核可配置向量核结构
- 1GHz@65nm，
- ~6亿晶体管，300mm²
- ~40W@1GHz
- 每秒万亿次操作（媒体操作）
- 双精度128GFLOPS
- 8000颗芯片构建千万亿次计算机
- 文章发表（录用）于Hotchip'10，ISSCC'11等





龙芯3B CPU在性能功耗比上达到了目前世界先进水平

芯片型号	频率 (GHz)	工艺 (nm)	核数	Die面积 (mm2)	功耗 (W)	双精度浮点峰值 (GFLOPS)	性能功耗比 (GFLOPS/W)
Intel Core i7 980 XE	3.2	32	6	240	130	107.55	0.827
Intel Sandy Bridge	3 ~ 4	28	8	370	130	256	1.96
AMD Opteron X12	2.4	45	12	346	130	152	1.16
IBM Power7	3~4.1	45	8	567	100	264.96	2.64
IBM PowerXCELL	3.2	45	9	221	80	100	1.25
Fujstu SPARC fxVIII	2.2	42	8	513	50	128	2.56
龙芯3B	1.0	65	8	300	40	128	3.2

为什么要加强计算机系统研究
— 从云计算谈起

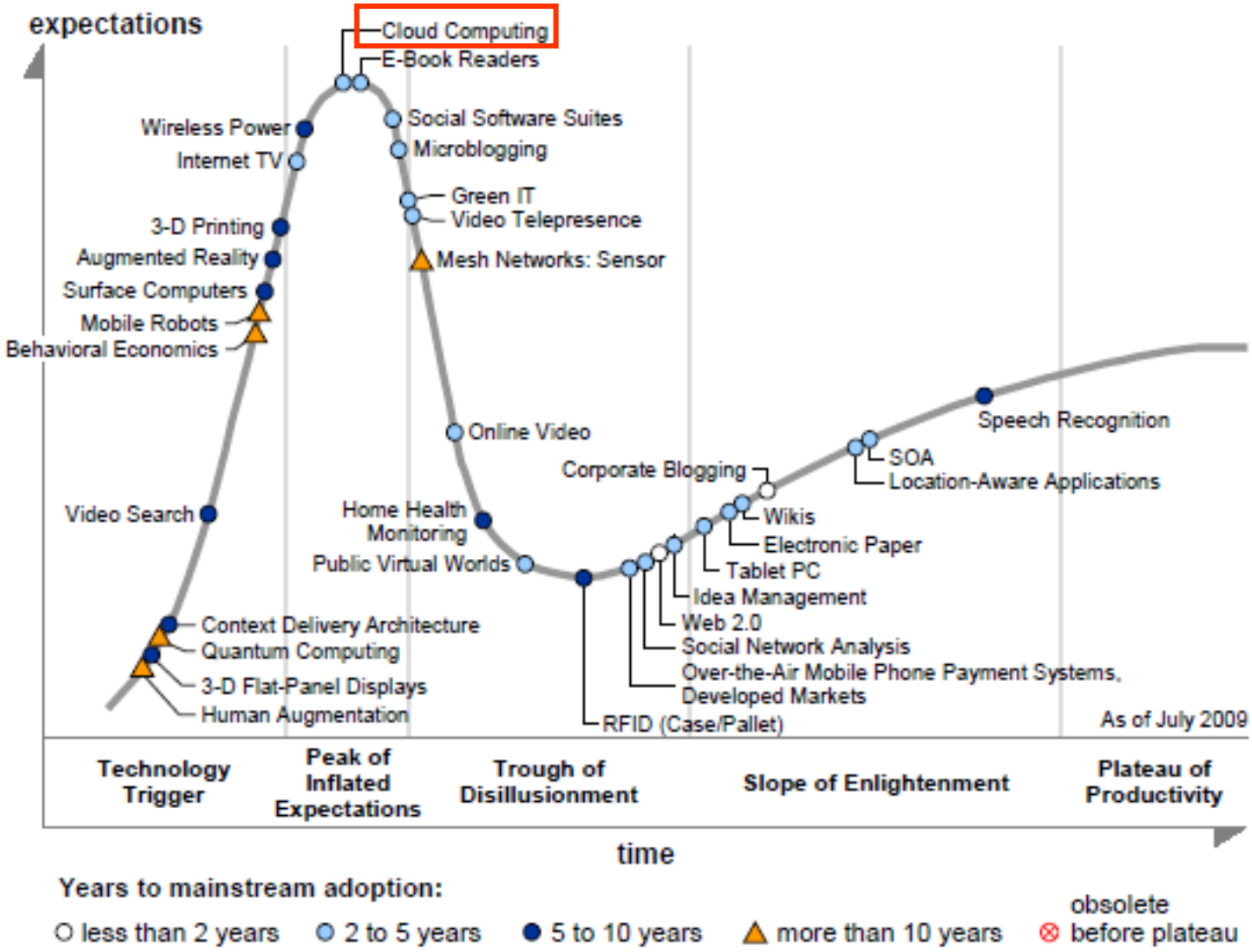
计算机系统技术面临转折性挑战

- 计算机系统结构（Architecture) 的概念是从上世纪60年代研制IBM360时提出来的，重要的贡献是区分了硬件与软件，定点与浮点，提出了系列计算机的概念，这些概念一直沿用至今。
- 现在和未来的云计算的workloads的特征与过去 HPC 、事务处理等应用有很大区别，从新的应用中（大量网络服务）应归纳出新的基本指令集。
- 今后决定一个云计算平台是否能存活不是光看虚拟化技术，而是看它的资源利用率，成本和可靠安全等系统因素。
- 认为计算机系统技术已经差不多了，国家可以不支持了，是一种非常缺乏远见的判断。把对云计算的支持偏重于中间件也是一种片面的政策。高效可靠的后台设备是云计算的关键。

网络问题逐渐变成计算机系统问题

- 电信业正在进入“后电信时代”。通信技术与业务正在趋向计算技术与应用；计算技术与应用正在趋向网络与服务提供，CT、IT正在真正走向融合。联通研究院将这种融合模式 称为“公众计算通信网（PCCN）“
- 在原有公众通信网的接入、交换、路由、传输要素的基础上，公众**计算**通信网还将实现计算处理能力、虚拟分配、调度管理以及业务开发等主要技术。
- 华为、中国移动等公司正在大量吸收懂系统结构的高端计算机人才。既懂计算机系统又懂通信协议的人才是目前最稀缺的人才。我国通信和计算机教育的分离不利于人才培养
- 通信领域两院院士**陈俊亮**教授到计算机学会担任**服务计算专业委员会主任**可看成一个标志性的事件。

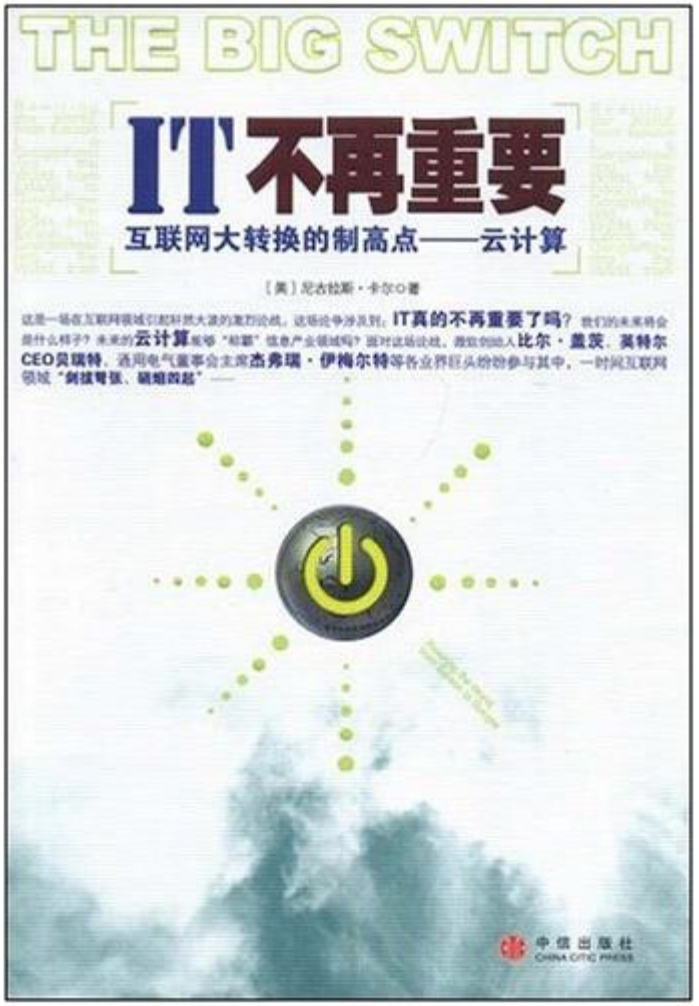
Emerging Technologies Hype Cycle 2009



Gartner

互联网不同于电网

- 对云计算的技术转换意义讲述最明白的书是Nicholas Carr写的“The Big Switch”,国内翻译成“IT不再重要”。
- 如同小型（直流）电站必然转变为大型（交流）电站一样，个人电脑必将让位于公共运算时代。
- 但云计算与电力网的不同，电力网只送能量不传送应用，所有的应用都是客户的责任；而信息服务应用可以通过网络传送。
- 电脑运算比发电更具模块性，数据存储、处理、传送可分拆成不同的服务。由不同公司提供，减少供应方的垄断。
- 云计算使万维网确实变成了万维电脑，成为我们感官和心灵的延伸。



云计算还不适合做尖端的超级计算

- Dan Reed：云计算绝对不是为特定目的构造的性能顶尖计算机的替代品。如果一种petascale计算需要极低的任务间通信延迟，今天的云计算肯定不适合。但是对于大多数使用较小规模设备的研究者，云计算是有吸引力的替代品.
- 目前的云模型并不支持顶尖的超级计算。动员grand challenge 应用的人做云计算就如同要说服驾驶第一方程式赛车的深受去乘公共汽车。
- HPC主要执行计算密集型的任务，CPU的利用率已经很高，虚拟化技术对提高HPC的CPU利用率作用不大。

目前的云计算做HPC效率较低

- 基于云计算理念来构建超级计算中心，除了满足传统的或现有的HPC用户需求外，更重要的是创造并吸引众多新领域的用户。
- 美国德州先进计算中心（TACC）的 Edward Walker 对Amazon EC2 上HPC应用的性能表现进行了研究，应用选择常用的基准测试程序NPB，测试结果表明：几乎相同的硬件条件下，对OpenMP版本的8个测试程序EC2性能下降7%至21%不等，MPI版本性能则下降40%至1000%不等。
- 虚拟化对计算密集型（如果数据能全部放进内存）应用的影响很小，而I/O密集型应用的性能则会有一定下降

在Amazon EC2 上运行MPI性能不高

- **Performance is below the level seen at dedicated, supercomputer centers, however, performance is comparable with low-cost cluster systems.**
- **Significant performance deficiency arises from messaging performance where latencies and bandwidths are between one and two orders of magnitude inferior to big computer center facilities.**

System	latency	uni-bw	bi-bw
LAM	81.20μs	57.85MB/s	81.98MB/s
GridMPI	83.46μs	54.60MB/s	77.07MB/s
MPICH2 nem	300μs	15.72MB/s	26.08MB/s
MPICH2 sock	85.87μs	58.49MB/s	83.42MB/s
OpenMPI	300μs	16.44MB/s	17.99MB/s
LAM/ACES	35.83μs	117.64MB/s	198.59MB/s

---MIT Constantinos Evangelinos and Chris N. Hill CCA-08 paper

目前推广云计算的重点在转变商务模式

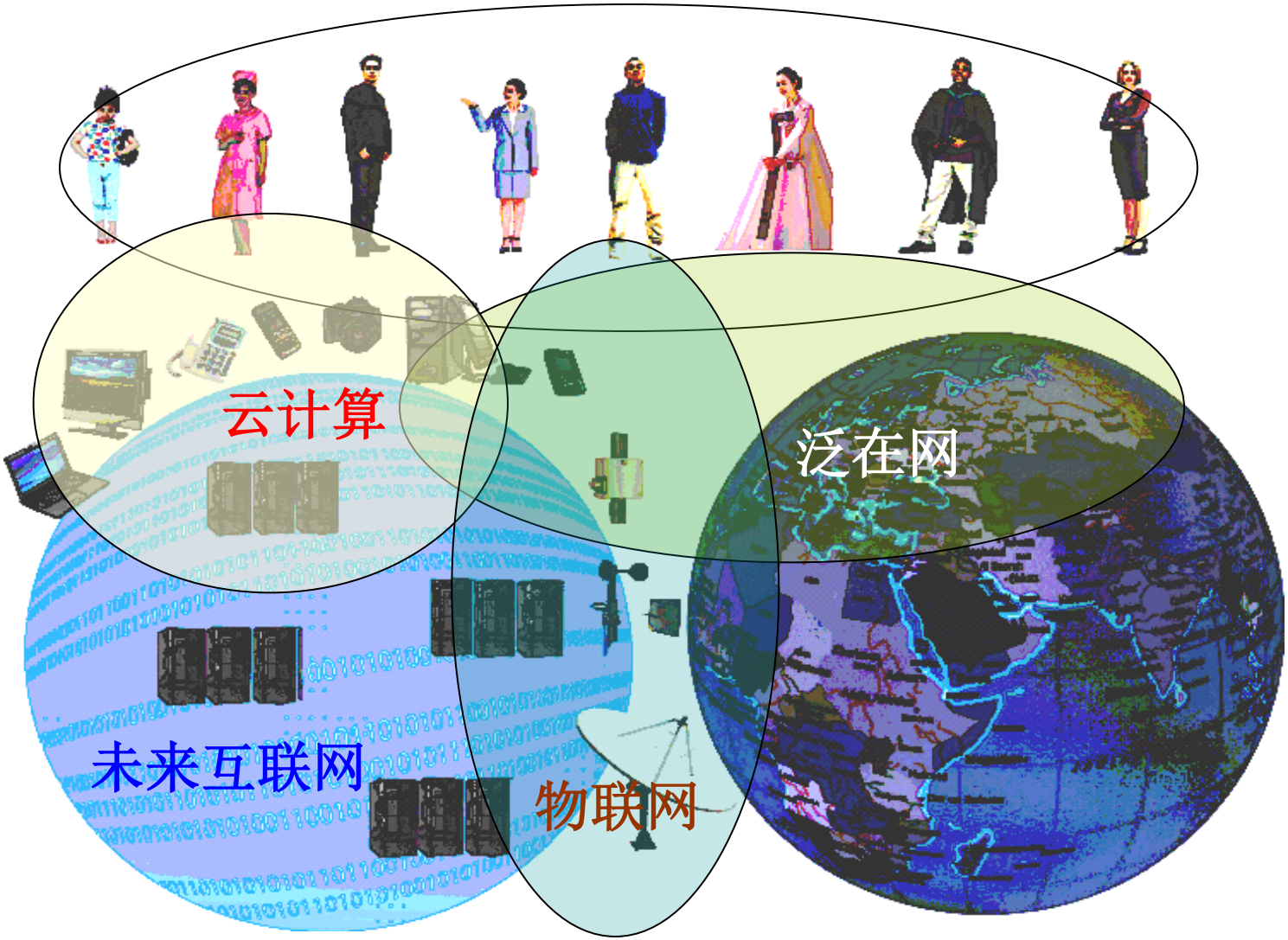
- 用户感觉到的云计算的好处主要是减少购买硬件软件的信息化开支，更好地满足动态变化的需求，降低用户端软件升级的维护成本和管理成本。这种好处主要来自**商务模式的转变**，其核心技术是**虚拟化技术**。
- 目前数据中心转向云计算平台的动力主要是**服务器的统计复用**，可以降低数据中心的运行成本，提高服务器和海量存储的**利用率**，其关键技术也是虚拟化技术。
- 虚拟化技术是一种相对门槛较低的技术，因此各大公司和各地政府都可以在较短时间内建立“云计算平台”。
- 实际上真正支持云计算的是**计算机系统技术**，这些技术用户看不见，媒体也很少宣传。
  - 与李开复的会面
  - 百度的看法

用户感觉不到的云计算技术

- 云计算系统的本质可以看成是：

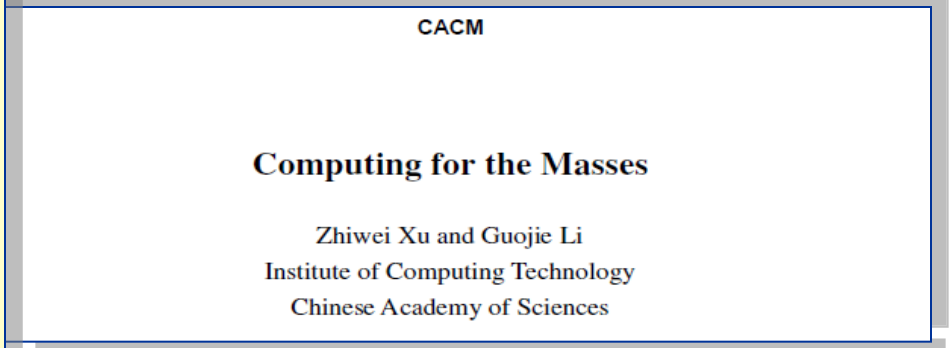
  资源虚拟化 + 并行计算
- 云计算不等于虚拟化。虚拟服务器并不能组成一朵云，云计算的能力远远超出一般的虚拟化解决方案。
- 并行技术是藏在云计算背后的核心技术，也是Google等云计算公司具有竞争力的关键技术。
- 各个层次的互连网络在数据中心起到一个非常核心的作用，其作用可能超过服务器本身。
- 虚拟化技术已经开始改变企业对服务器、操作系统以及计算资源的重新部署，甚至导致全新的IT管理模式，这一变化无疑将对操作系统产生重大影响。

未来互联网、物联网、泛在网和云计算



“Computing for Masses” 是计算机领域的一场技术变革

- Computer Science for the Ternary Universe
  - Algorithm Networks
  - Ternary Systems Modularity
  - Fundamental Impossibility
- A Universal Compute Account for Everyone
- Lean System Platforms for the Masses
- A Science of the Net Ecosystem
- National Information Accounts


CACM

Computing for the Masses

Zhiwei Xu and Guojie Li
Institute of Computing Technology
Chinese Academy of Sciences

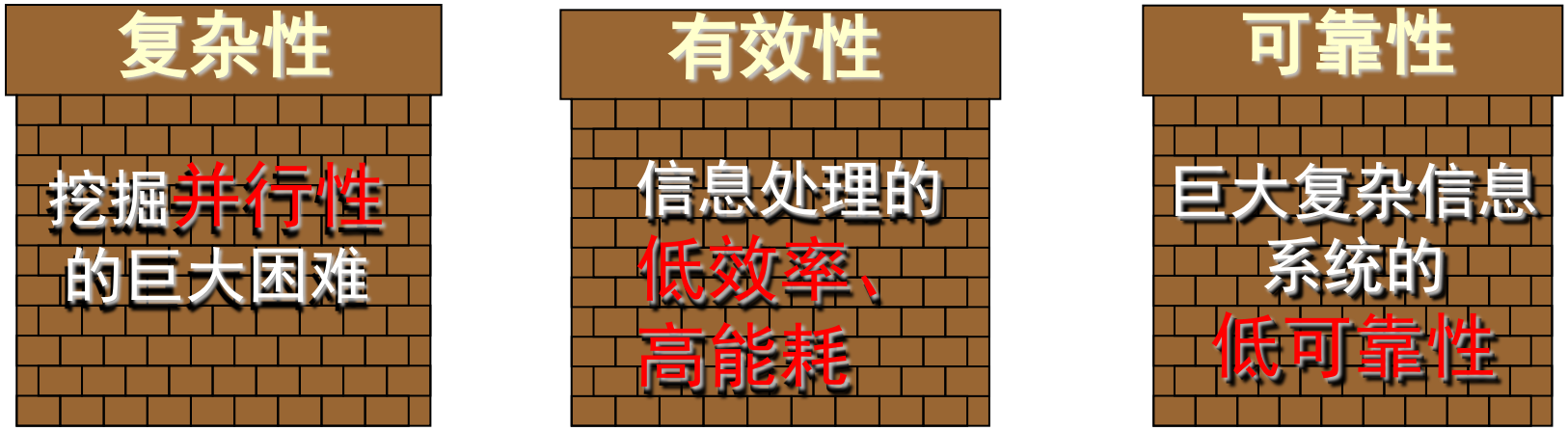
- **Computing for the masses means providing essential computing value for all people, tailored to their individual needs.**
- **It demands paradigm-shifting research and discipline rejuvenation in computer science, to create augmented Value (V), Affordability (A) and Sustainability (S) through Ternary computing (T).**
- **Computing for the masses is VAST computing.**

获图灵奖的与系统有关的计算机科学家

- 2009 Thacker, Charles P Alto personal computer, Ethernet .
- 2008 Liskov, Barbara fault tolerance, and distributed computing.
- 2007 Clarke, Edmund M，Emerson, E Allen， Sifakis, Joseph
  effective verification technology
- 2006 Allen, Frances optimizing compiler techniques
- 2005 Naur, Peter ALGOL 60 compiler
- 2004 Cerf, Vinton，Kahn, Robert E internetworking, TCP/IP
- 2002 Adleman, Leonard M. Rivest, Ronald L，Shamir, Adi
- public-key cryptography
- 1997 Douglas Engelbart mouse GUI
- 1992 Bulter Lumpson PC environment
- 1990 Fernando J. Corbato 资源共享的计算机系统开发
- 1987 John Cocke 开发RISC计算机
- 1983 Ken Thompson、Dennis M. Ritchie UNIX操作系统。
- 1967 Maurice V. Wilkes 存储程序的计算机
- 1966 A.J. Perlis 先进编程技术和编译架构

计算机系统研究
需要突破的关键技术

计算机系统难以逾越的三座高墙



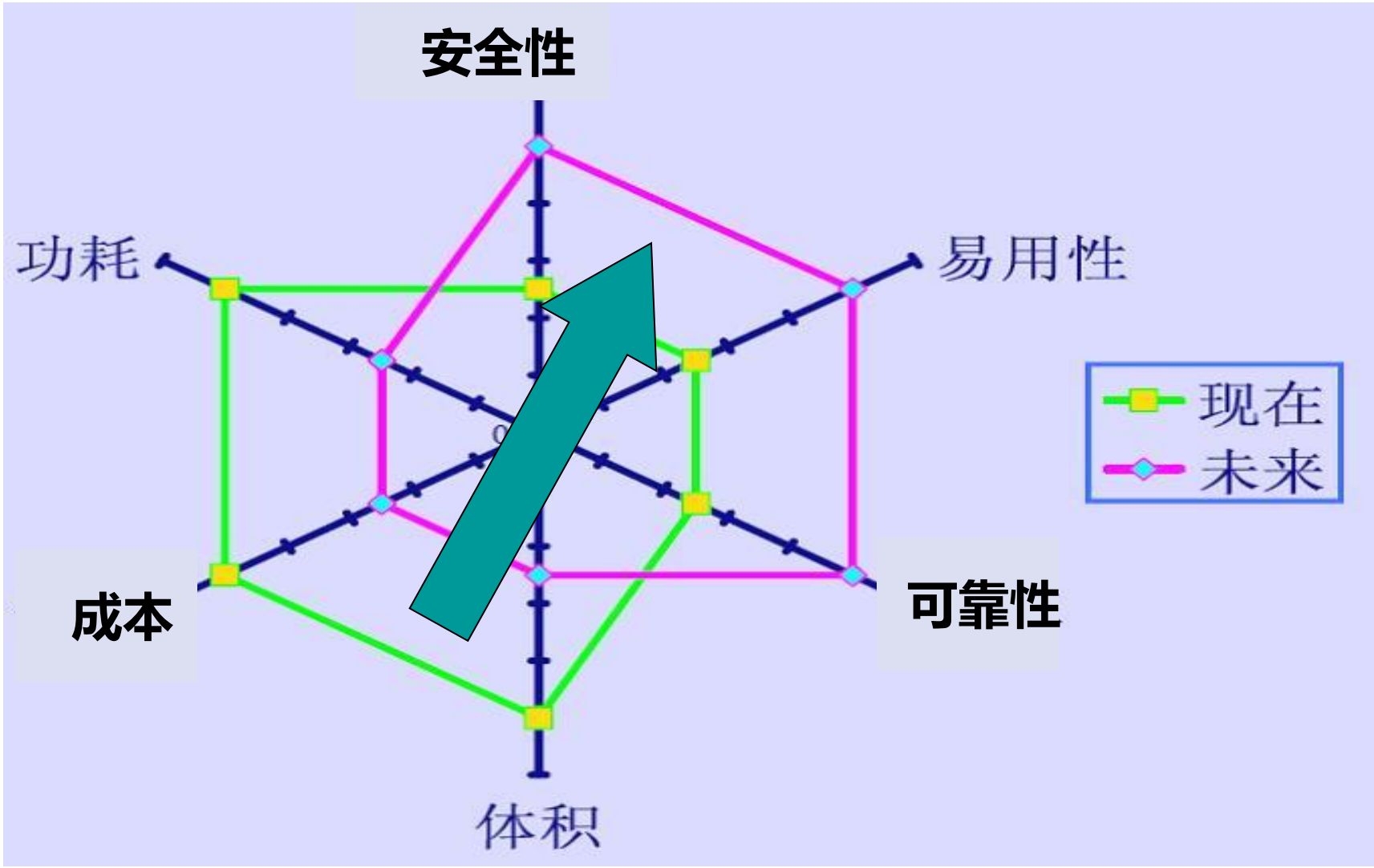
- 到2020年前后，硅技术物理极限临近，摩尔定律将终止，计算机系统的发展将遇到难以逾越的巨大技术障碍，必须有重大理论和结构上的突破
- 新型的体系结构方向的研究必须展开，以突破并行、功耗和可靠三座技术高墙

计算机系统研究的基本问题

- 如果你问Patterson等国外学者:”你是研究什么的？“他们一般会简洁地回答：”System”。我国计算机学者一般很少讲自己是研究系统的。研究计算机系统的学者关心什么？
- 计算机系统研究关注的主要问题包括：
  - 计算机指令系统
  - 适应各种应用的系统结构（通用与专用）
  - 资源利用的效率（包括虚拟化技术）
  - 计算机使用的方便性和灵活性
  - 编程的效率(尤其是并行变成）
  - 计算机的性能和可扩展性
  - 系统的可靠性与安全性
  - 降低计算机生命周期的总成本
  - 减少计算机的能耗

目前的云计算只涉及其中少数问题，许多基本问题有待解决。

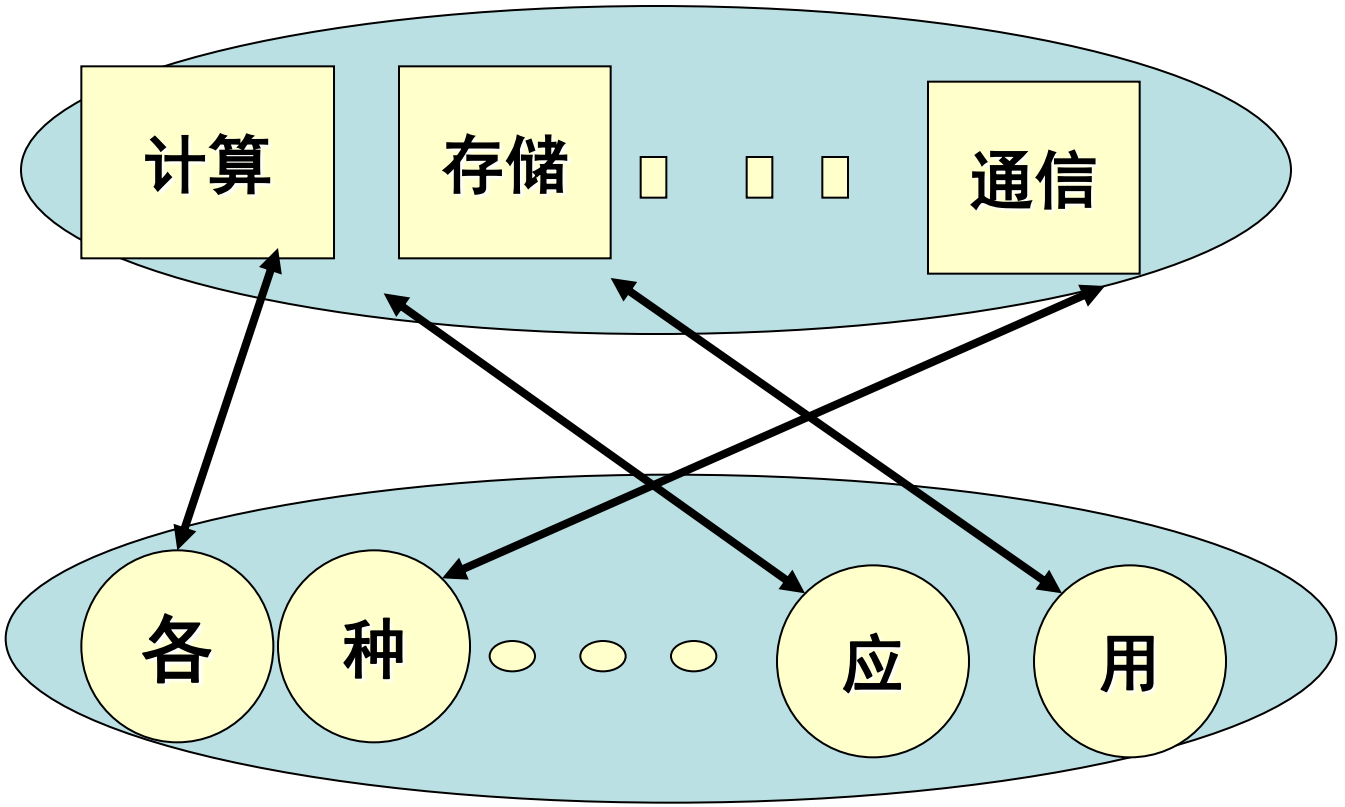
改变计算机系统技术的研究方向



计算机应用的“昆虫悖论”

- 用一句话概括21世纪信息技术的发展趋势就是“为大众计算(Computing for the Masses”)。
- 数十亿用户和各行各业的应用需求一定千差万别。日本东京大学的坂村健教授曾把PC机、手机和物联网的应用种类分别比喻成哺乳类动物(2万种）、鱼类（3万种）和昆虫类（100万种）。计算机系统如何满足如此多的应用。
- 为每一种应用设计一种专用芯片和系统对供应商不经济，采用同一种通用计算机对用户而言效率不高。通用和专用是计算机系统发展中永恒的矛盾，也是最大的挑战。
- 可重构芯片可计算机，可复用的软硬件模块是计算机系统研究的追求目标。虚拟化技术也是解决此矛盾的途径之一

计算机系统研究最基本的问题——满足应用需求的计算资源配置



云计算是MRMT系统

- 我们可以仿照Flynn的计算机分类，将计算机系统按资源自治域（Resource)－任务(Task）分成4类（资源自治域的概念是借用互联网自治域的提法，需要认真定义和研究）
  - SRST（单资源单任务系统）1对1 , 如低端手机
  - SRMT（单资源多任务系统）1对多，如mainframe
  - MRST（多资源单任务系统）多对1，如某些HPC
  - MRMT（多资源多任务系统）多对多, 如云计算
- 人们常说网格是多对1的系统，云计算是1对多的系统。实际上云计算的后台有许许多多资源，很难在一个操作系统控制之下，是一个典型的多对多的系统。
- 数据中心的庞大的计算、存储资源如何高效的调配是计算机系统研究的大问题。关键是资源是不是在统一的调度系统管理之下。

计算机系统的难点在并行处理

- 并行处理已研究了几十年，论文多如牛毛，但进展不大。
- 云计算号称用1000台服务器工作1小时的成本与用一台服务器工作1000小时相当。问题是效率怎样，如果只完成了单服务器的1/10工作，仍然不合算。
- 并行计算最关心的“如何提高计算机的性能和效率”，这个问题从来没有改变，但答案在不断变化。
- 影响并行效率的障碍一是编程（人工效率），二是通信延迟与带宽。“带宽墙”可能比“存储墙“和“功耗墙”更高。
- 历史上计算机设计的匹配规律是完成一次浮点运算需要保证一个字节的供数能力。目前主流CPU的运算速度与供数带宽之比是1：0.3-0.5，即100Gflops的芯片需要50GBps左右的内存带宽（4~8个DDR3)。GPU芯片一个字节供数要完成十次以上浮点运算，典型的“茶壶煮饺子”。

挖掘并行性是计算机系统的巨大挑战

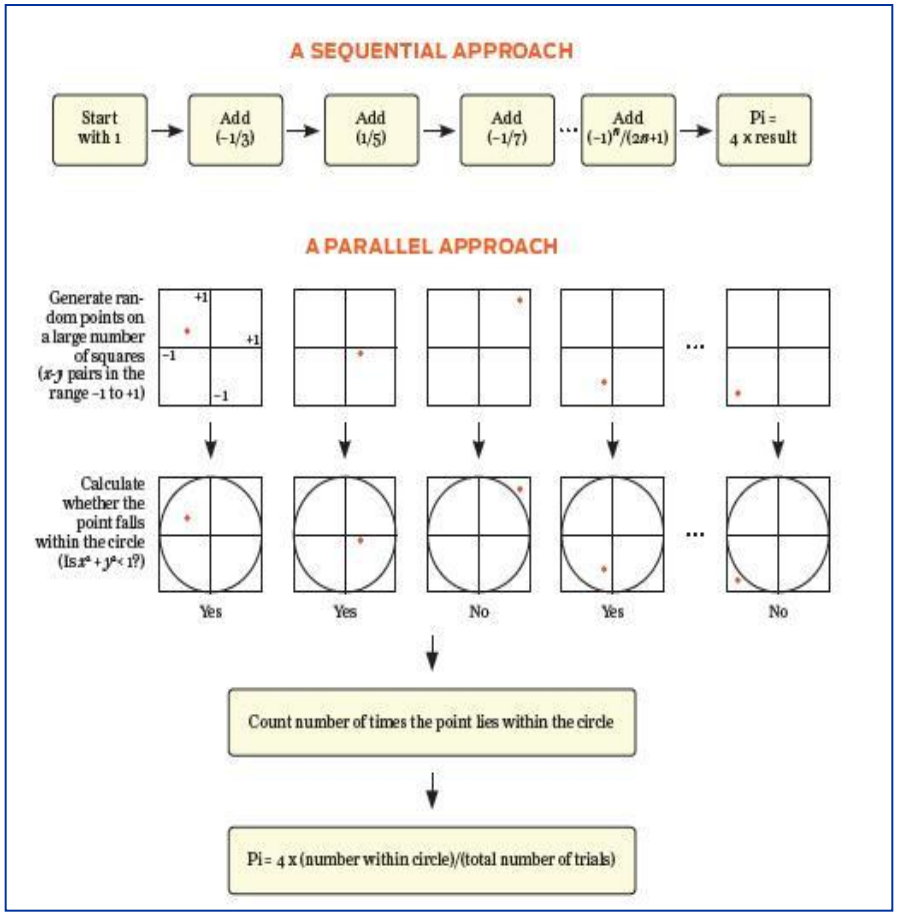
时间	2020年	2030年	2050年
器件	CMOS	纳米量子器件	量子、生物分子
计算速度	Exaflops (10^{18})	Zettaflops (10^{21})	>Yottaflops (10^{24})
并行度	10^{8-9}	10^{10} — 10^{12}	10^{13}— 10^{15}
内存容量	25PB	EB (10^{18}B)	ZB (10^{21}B)
功耗	40MW	MW	MW
用途	核聚变模拟 蛋白质折叠等	地球模拟 生命科学等	MEMS优化 脑科学模拟等

2010 2020 2030 2050

设计并行算法往往要“从头开始”

- 直接将现有串行程序并行化效率不高，需要从原始问题开始。
- 已提出数百种并行编程语言，例如APL, Id, Linda, Occam, and SISAL，至今无一成功。
- 自动并行的成功率与CPU核的数目乘反比。
- 容易并行的应用包括
  - 任务级并行
  - 数据并行（如GPU）
  - 某些科学计算

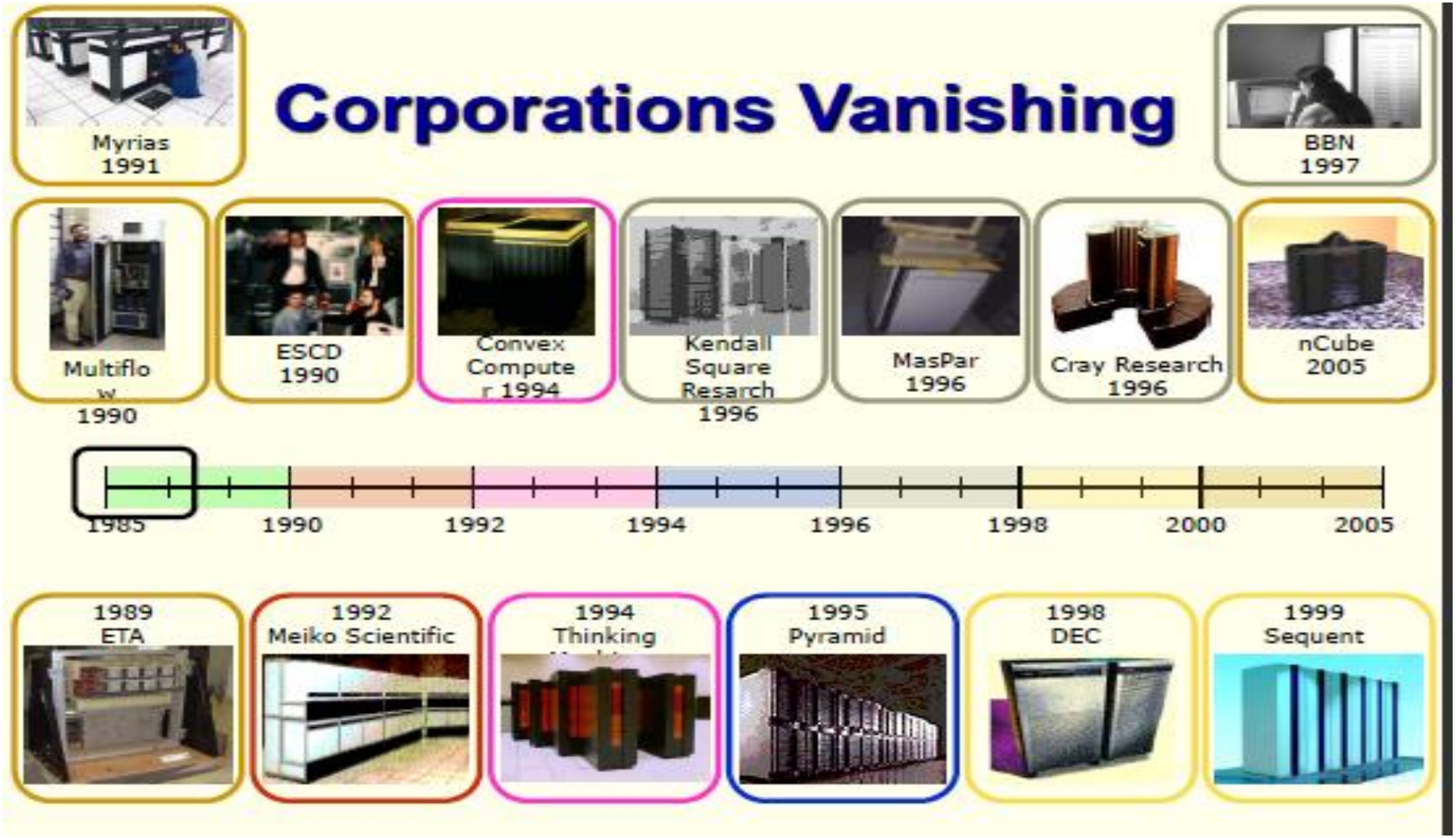
$PI=4*\sum(-1)^n/(2n+1)=4*(1-1/3+1/5-1/7+1/9.....)$



这一场“并行革命” 可能失败

- **人生三件很不愿做又不得不做的事：纳税、死亡，并行处理！**
- 2007年1月， Stanford大学校长，计算机体系结构领域的权威学者 John Hennessy在ACM杂志上指出 :“当我们谈论并行性和轻松地使用真正的并行计算机时，我们是在谈论一个计算机科学家面对的最困难的问题，如果我在计算机企业，我将感到恐慌。”
- IT产业从一个高成长的产业变成一个等待替代产品的产业，我们怎么办？如果软件不能有效地利用几十甚至上千个片内CPU核，计算机就不可能更新换代了，这是一个巨大的危机。

被扔进历史垃圾桶的并行计算机

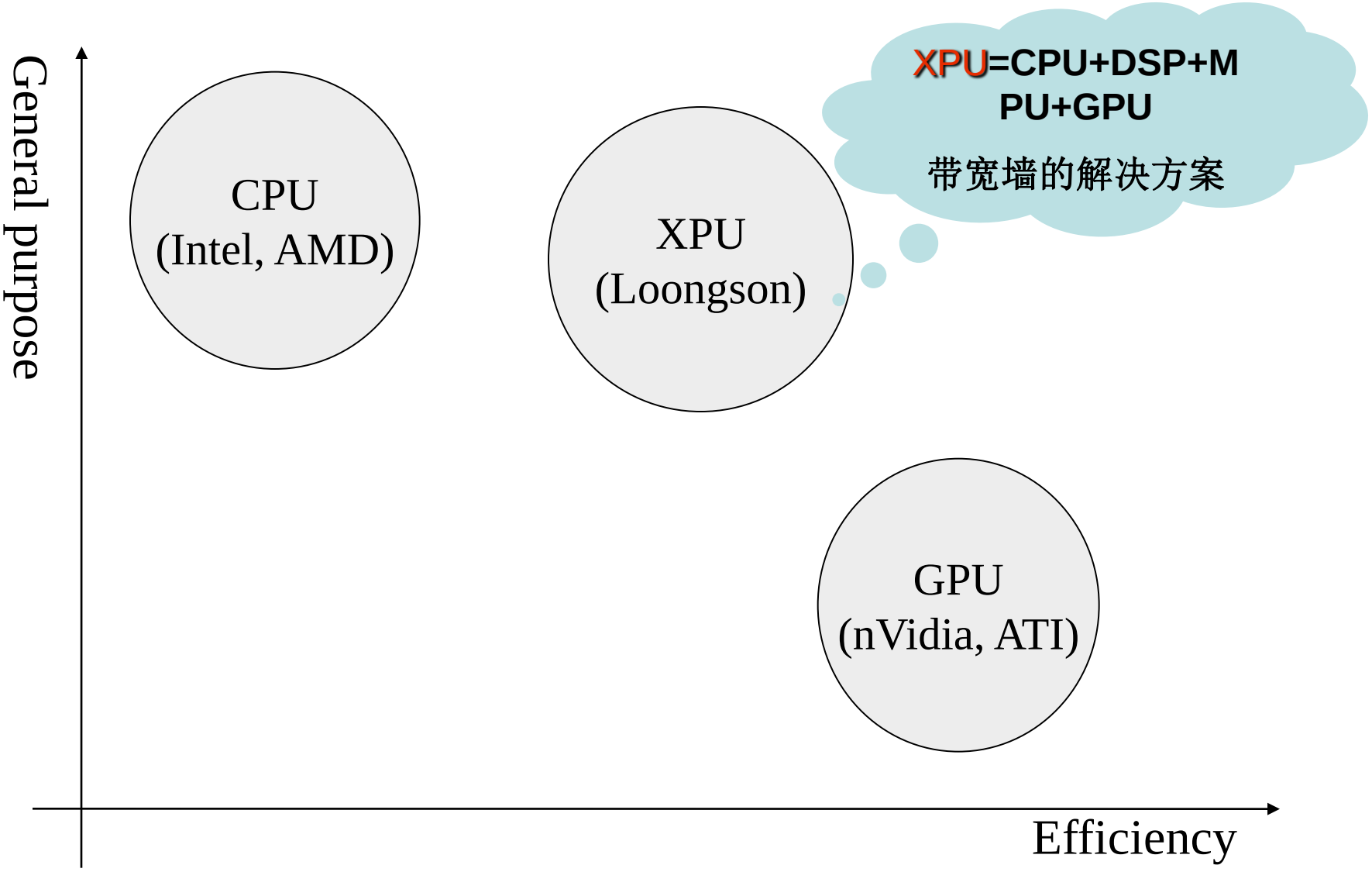


Multicore Is Bad News For Supercomputers

- "The Troubles with Multicores", David Patterson, IEEE Spectrum, July, 2010
- "球"已扔出，谁来接球?

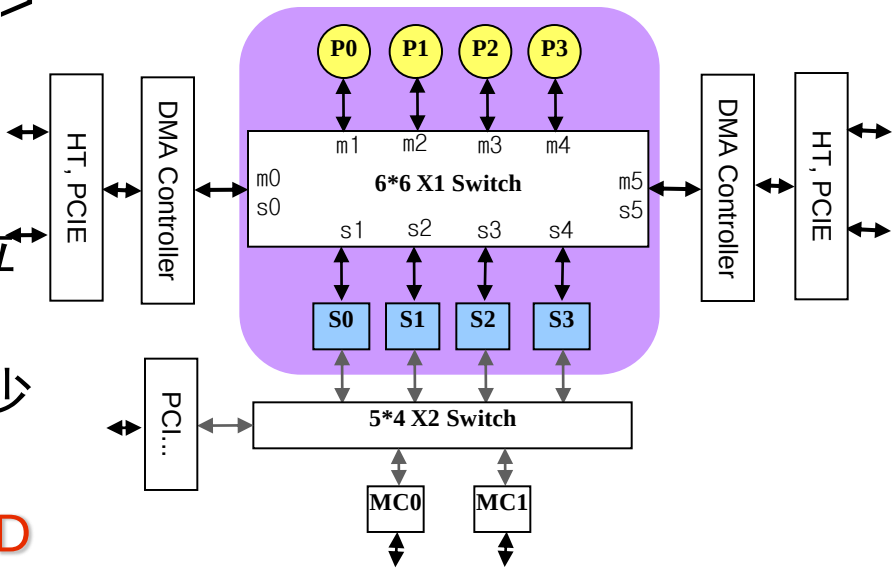
- Throughput = Concurrency/Latency
  - Exploiting parallelism
  - Exploiting locality
- Multi-core Cannot Deliver Expected Performance as It Scales
- "Multicore Is Bad News For Supercomputers", Samuel K. Moore, IEEE Spectrum, Nov, 2008
- Memory wall
- Bandwidth wall

XPU处理器核体系结构



I/O为中心的处理器体系结构

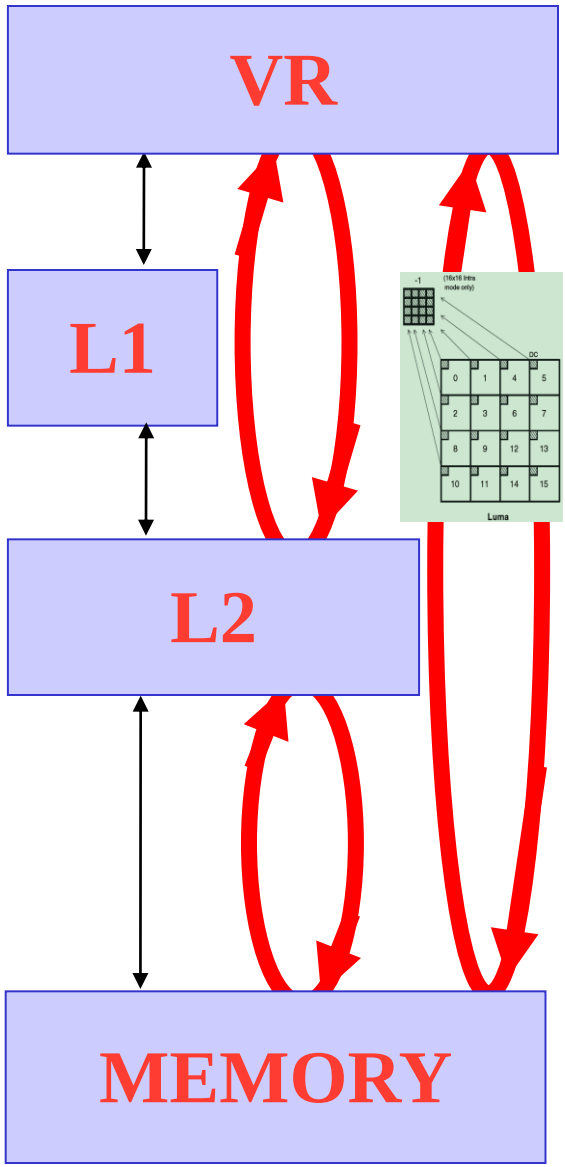
- 以互联网络应用为代表的高吞吐率计算成为主流
- 处理器设计阶段：计算为中心->存储为中心->I/O为中心
- 面向I/O为中心的结构
  - 提升I/O系统在处理器存储层次中的位置，使I/O系统高于处理器二级cache
  - 优点：能够利用片上二级cache来减少对内存访问，提高性能
  - 应用于龙芯3号设计，测试表明：SSD磁盘访问性能提高了40%，相对Intel平台有15%的性能提高
- 相关文章发表于IEEE Micro 2009、HPCA'2010



图：四核龙芯3号结构

龙芯XPU 体系结构-GS464V

- 功能强大的向量运算部件
  - 每个核两个256位向量部件
  - 新增300多条SIMD 指令 (Linpack, FFT, filter, media……)
- 专门的数据通道
  - VR、L2、MEM之间的专门链路，类似DMA
  - 运算和数据引擎并行
  - 数据传送过程中的重构：矩阵转置、FFT位反、媒体墒解码….
- 融合通用性与运算效率
  - Linpack效率>90%,
  - FFT效率>85%,
  - 单核1080p的H.264解码>100帧/秒
- 在HotChips'2010上发表



计算机系统必须解决存储层次问题

• L1 cache reference:	0.5 ns
• Branch mis-predict:	5 ns
• L2 cache reference:	7 ns
• Mutexlock/unlock:	25 ns
• Main memory reference	100 ns
• Compress 1K Bytes with Zippy	3000 ns
• Send 2K Bytes over 1 GBPS network	20000 ns
• Read 1 MB sequentially from memory	250000 ns
• Round trip within data center	500000 ns
• Disk seek	1000000 ns
• Read 1MB sequentially from disk	2000000 ns
• Send one packet from CA to Europe	15000000 ns

“互连为王”— 数据中心现状

互连带宽和延迟是必须啃下的硬骨头

- 未来的数据中心需要与目前主流服务器不同的计算机系统来满足云计算的要求。
- 数据中心需要具有大量接口的网络交换器，其价格远远高于市面上流行的交换器，比普通交换器对分带宽bi-section bandwidth）高10倍的交换器的价格要高出100倍。
- 目前的以太网技术无法让数据的传输速率超过每秒100G，主要因为没有这么多的能量来给提供这种数据传输速度的网络系统提供电力和进行冷却。
- 3D芯片和使用硅基光电子学来制造低成本、集成、传输速率达TB级的互连是解决互连问题的希望

降低系统功耗的多种途径
—途径非常多意味着没有找到真正的途径

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]

信息为什么这么“重”？
—解决计算机功耗问题的联想

- 假设要传送200TB的数据从北京到西安（1200公里），按200GB的硬盘一公斤计算，大约一吨重，按目前货运市价每吨公里）0.1元左右计算，运费大约120元，加上其他开销不会超过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。
- 这就相当于火车运送硬盘的“原子”只有1吨重，但需要送过去的“BIT”按运费算超过100吨，BIT比原子“重”100倍。
- 如果把全国的网络带宽增加100倍，即从10Gb到1Tb, 可能全国发的电一半要用在网络上（现在占5%左右）

Google 达拉斯数据中心

- 占用了附近一个180万千瓦（长江三峡发电站的1/10）水力发电站的大部分电力输出，利用河水冷却服务器。

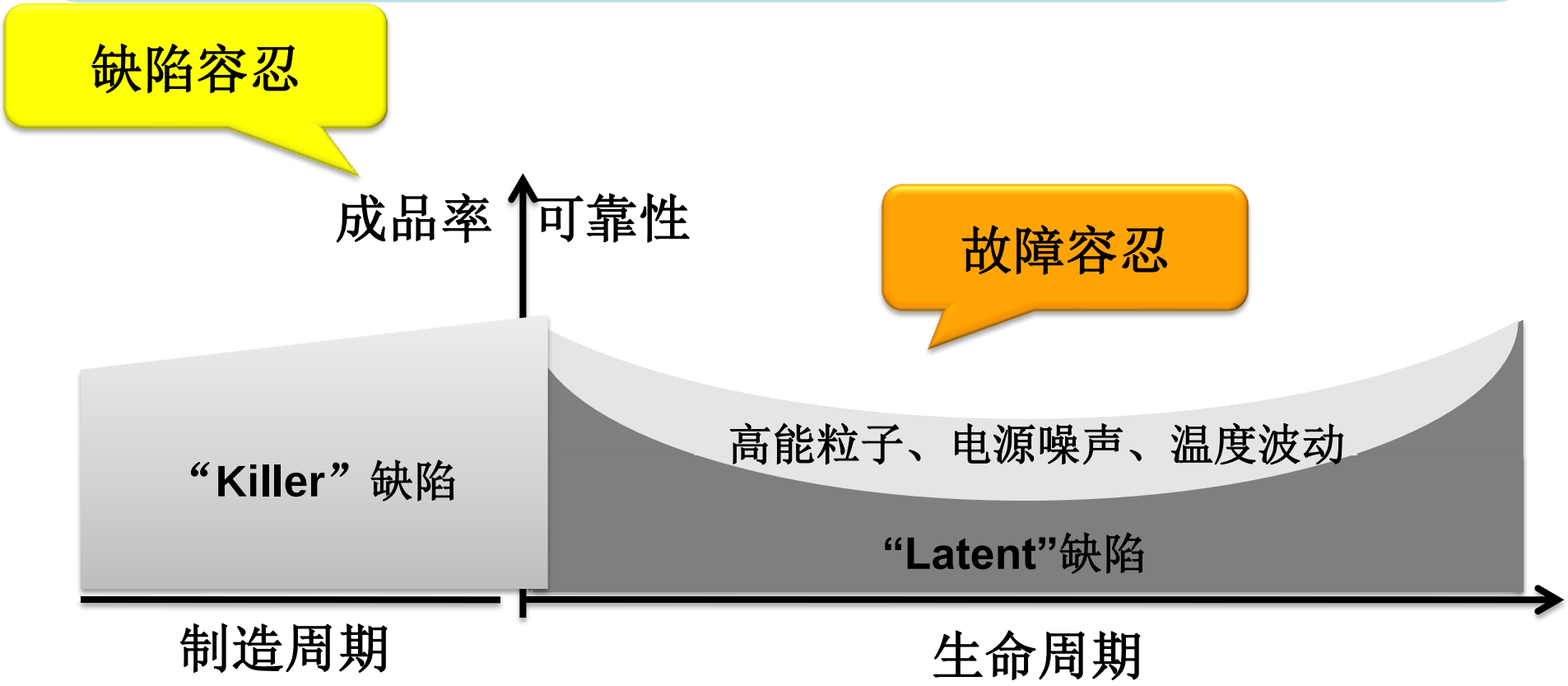
网上公开的Google技术

- Google的每个服务器机架内部连接每台服务器之间网络是100M以太网，在服务器机架之间连接的网络是1000M以太网。
- 目前 Google 已经在全球运行了38 个大型的IDC中心，超过300 多个GFSII 服务器集群，超过80万台计算机。
- Google 所拥有的八十万台服务器都是自己设计打造的，Google 认为这是公司的核心技术之一。
- 每个服务器刀片自带12V 的电池来保证在短期没有外部电源的时候运行

可靠性设计面对的科学问题

自测试自诊断自修复—3S原理

如何在由数十亿个元件组成的芯片上构造稳定可靠的系统？



改变不触动“核心技术”的科研模式

请批评指正！