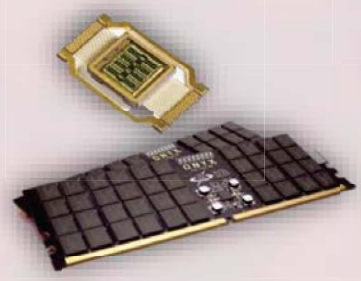
新型存储技术进一步推进内存数据库发展

- 更大的存储密度
- 更低廉的价格
- 更低的能耗
- 字节级寻址能力
- 更好的随机写性能
- 更好的持久存储能力

- 非易失存储器（Non-Volatile Memory，NVM）
- NVM主要类型：
  - 相变存储器（Phase Change Memory, PCM）
  - 磁阻式存储器（Magnetoresistive Random-Access Memory，MRAM）
  - 阻变式存储器（Resistive Random Access Memory，RRAM）
  - 铁电存储器（Ferroelectronic RAM, FeRAM）
  - Intel 3D Xpoint

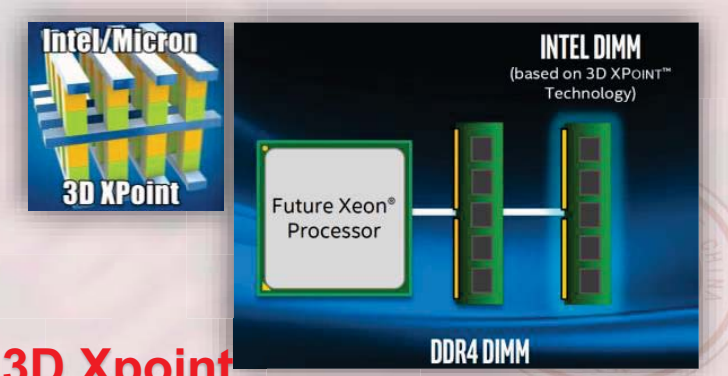


PCM

MRAM

FRAM

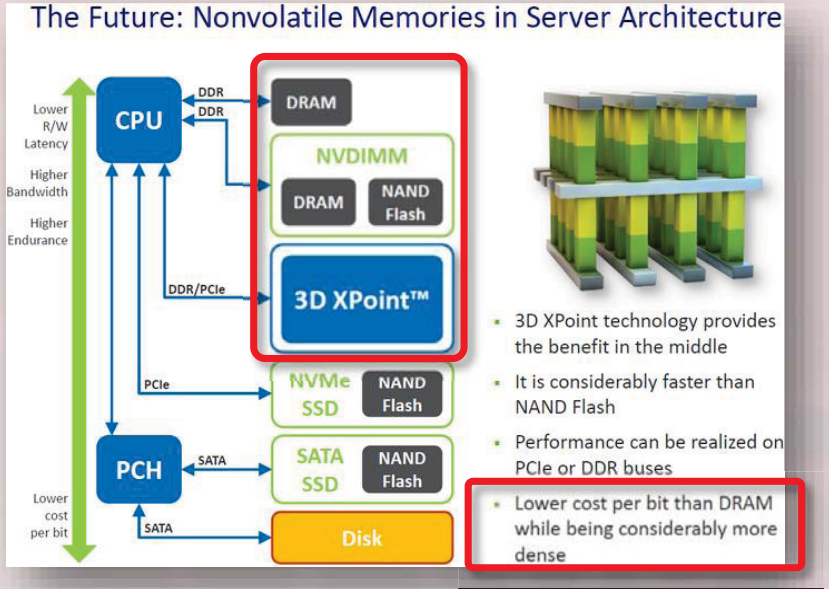
RRAM

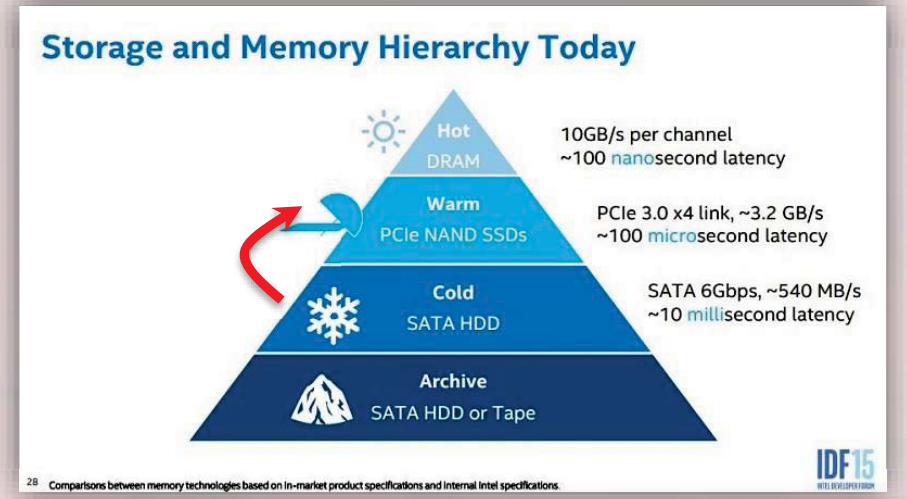


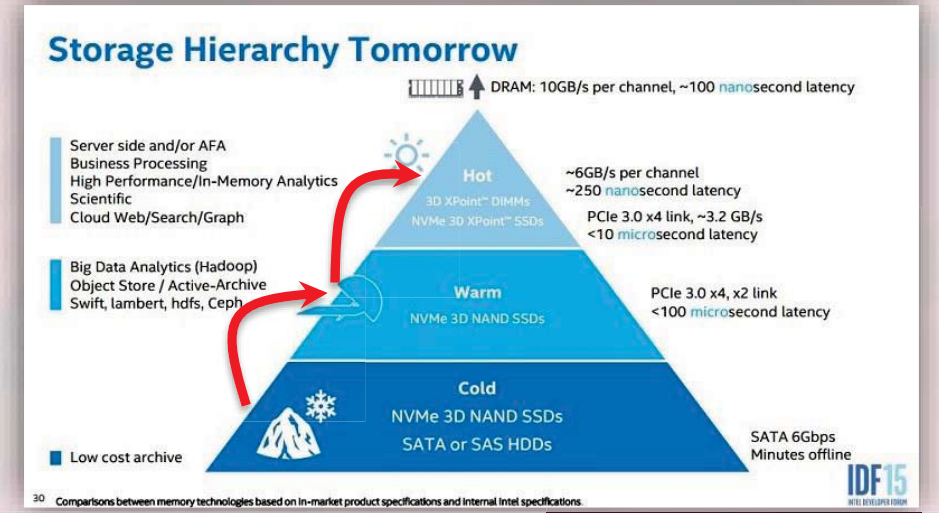
3D Xpoint

新型存储改变存储层次设计

- 未来的3D Xpoint非易失性内存可能会逐渐取代DRAM成为持久化热数据存储层
- 新型非易失性内存存储密度更大，成本更低，支持更高性价比的内存计算应用
- 新型存储简化存储层次，推动数据库存储引擎升级

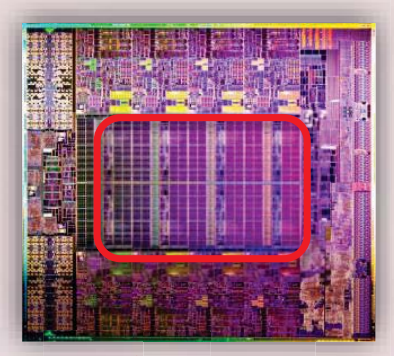




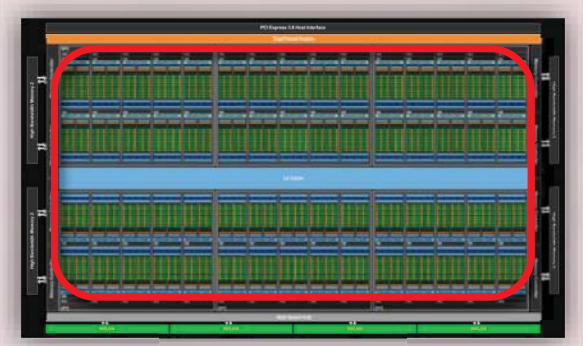


处理器技术进步推动内存数据库发展

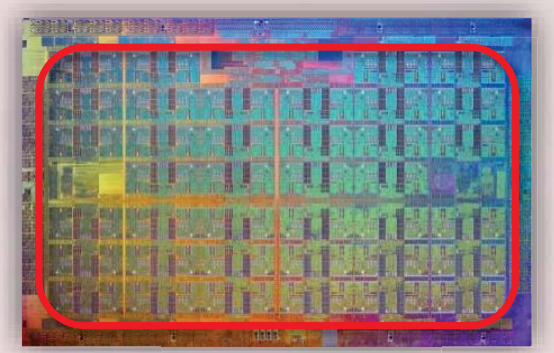
- 处理器从多核走向众核，从同构走向异构
- 查询优化技术从以cache为中心走向以高并行计算为中心



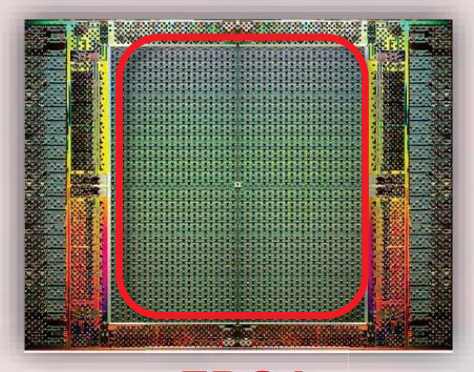
多核CPU



GPGPU

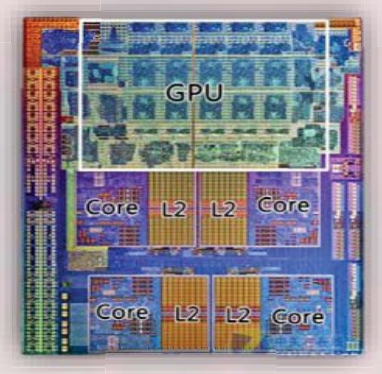


Xeon Phi

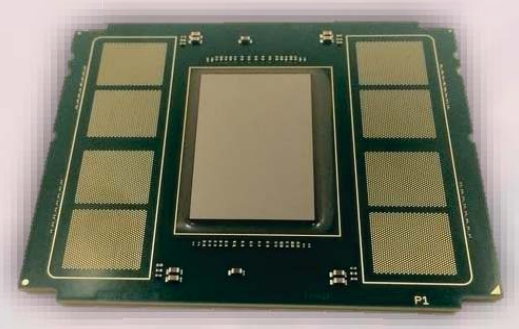


FPGA

融核架构



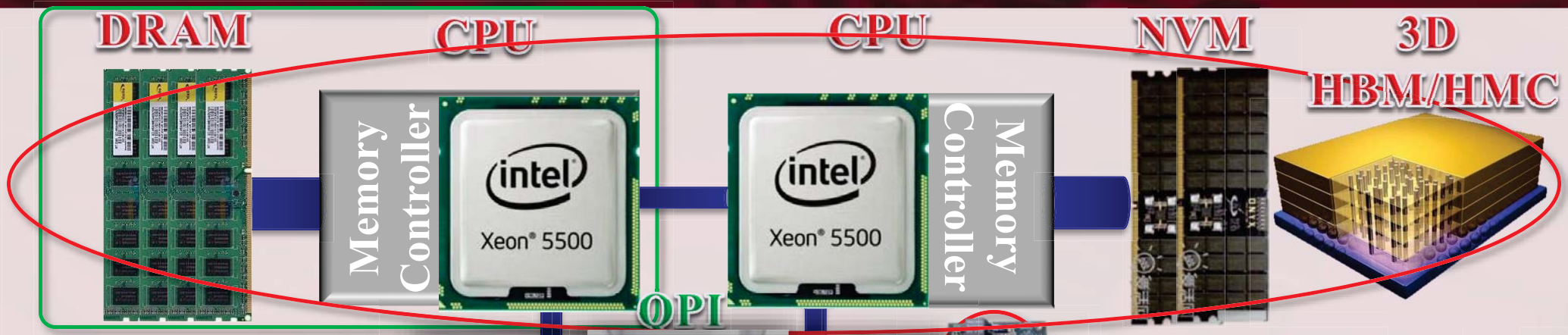
APU

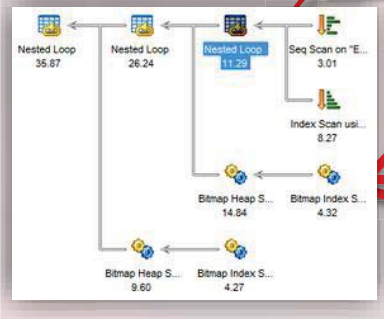


KNL Phi

Xeon + FPGA

内存数据库技术挑战——异构内存计算架构



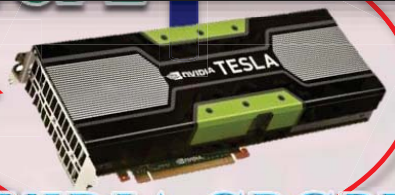


Intel Xeon Ph

FPGA

CPU与协处理器如何协同

协处理器如何实现查询处理



NVIDIA GPGPU

如何消除PCIe数据传输瓶颈

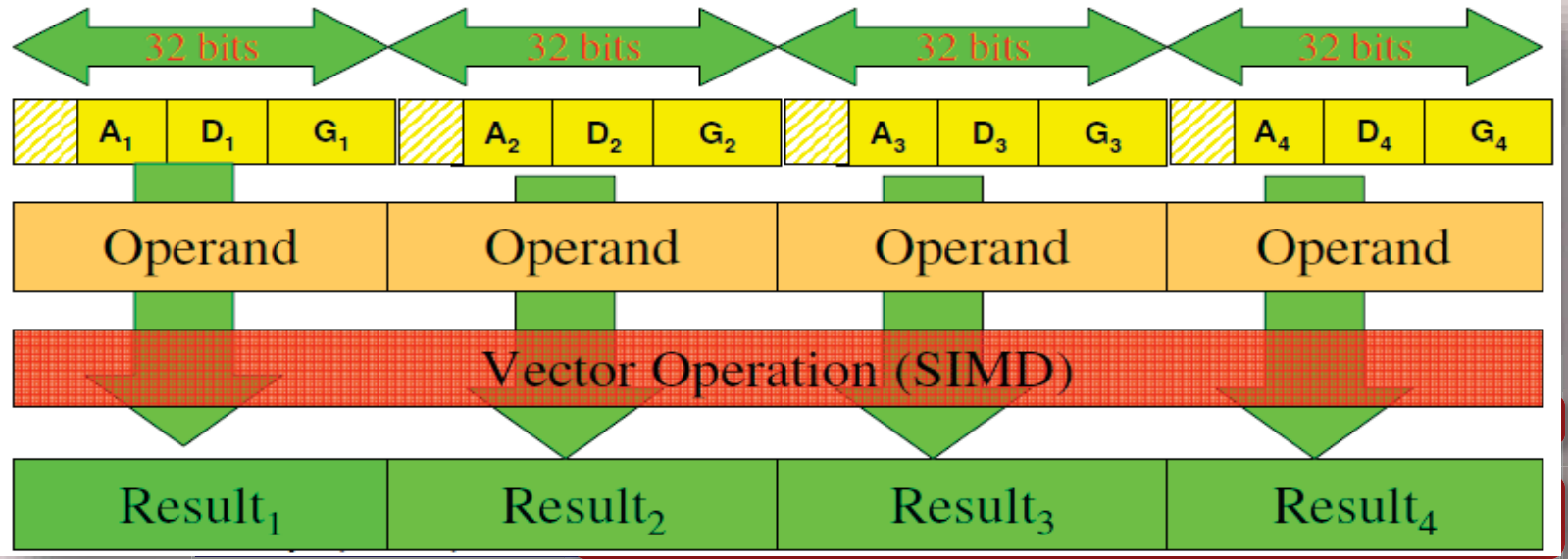
代表性多核CPU与众核协处理器

- 从多核CPU到众核处理器
  - 由并行处理到高并行处理：从几十到几千并行处理线程
  - 从128位SIMD向量处理到512位SIMD处理
    - 通过SIMD提高排序及哈希连接等数据库操作性能
  - 从较小的cache到GB级设备内存
    - 更高带宽：~100GB/s vs. 300+GB/s
    - 存储访问优化：从以cache优化为中心到高并发内存访问优化

类型	Xeon E7-8890V3	Xeon Phi 7120X	NVIDA Tesla K80
核心数量/线程数量	18/36	61/244	4992 CUDA cores
主频	2.50 GHz	1.24 GHz	562MHz
内存容量	1.54 TB	16GB	24GB
缓存容量	45MB	30.5 MB	--
内存类型	DDR-4	DDR-4	GDDR5
内存带宽	102GB/s	352 GB/s	480 GB/s
价格	$7174.00	$4129	$3,495.00

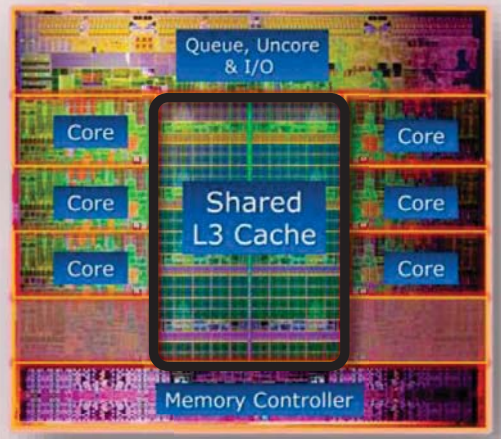
代表性多核CPU与众核协处理器

- 从多核CPU到众核处理器
  - 由并行处理到高并行处理：从几十到几千并行处理线程
  - 从128位SIMD向量处理到512位SIMD处理
    - 通过SIMD提高排序及哈希连接等数据库操作性能



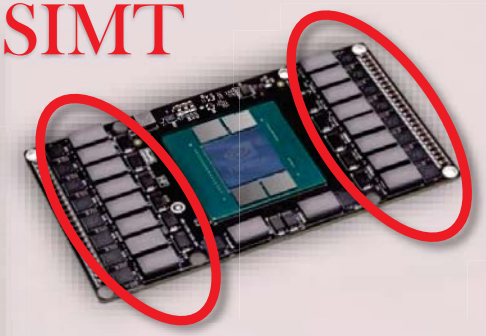
内存类型	DDR-4	DDR-4	GDDR5
内存带宽	102GB/s	352 GB/s	480 GB/s
价格	$7174.00	$4129	$3,495.00

多计算平台算法优化策略



Caching

SIMT



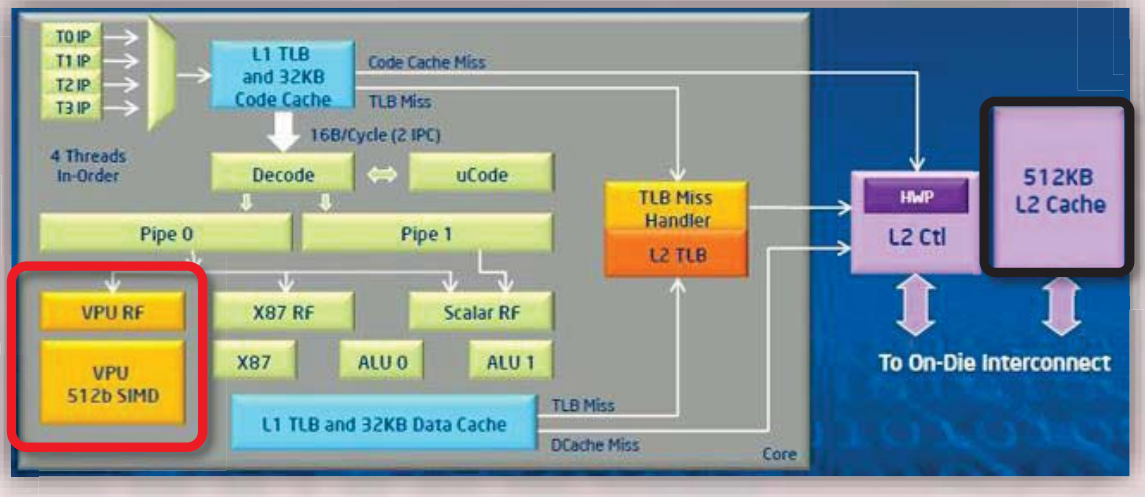
On-board memory

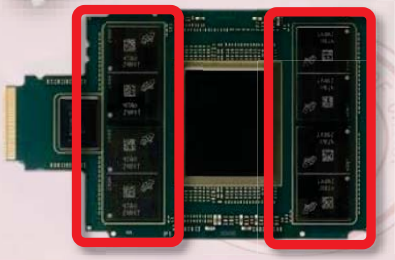
SIMD

Massive

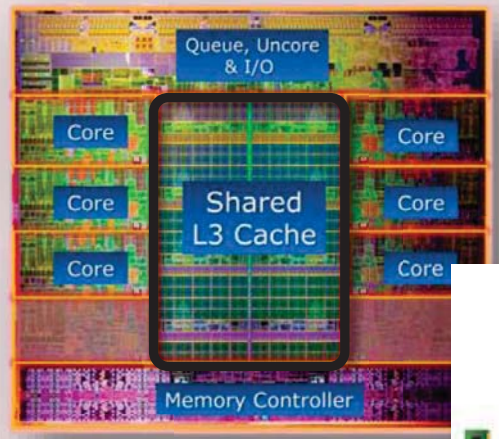
Parallel

Threading





多计算平台算法优化策略



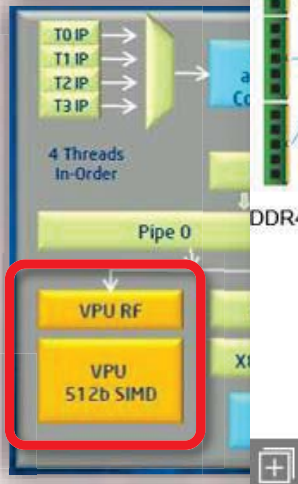
Caching

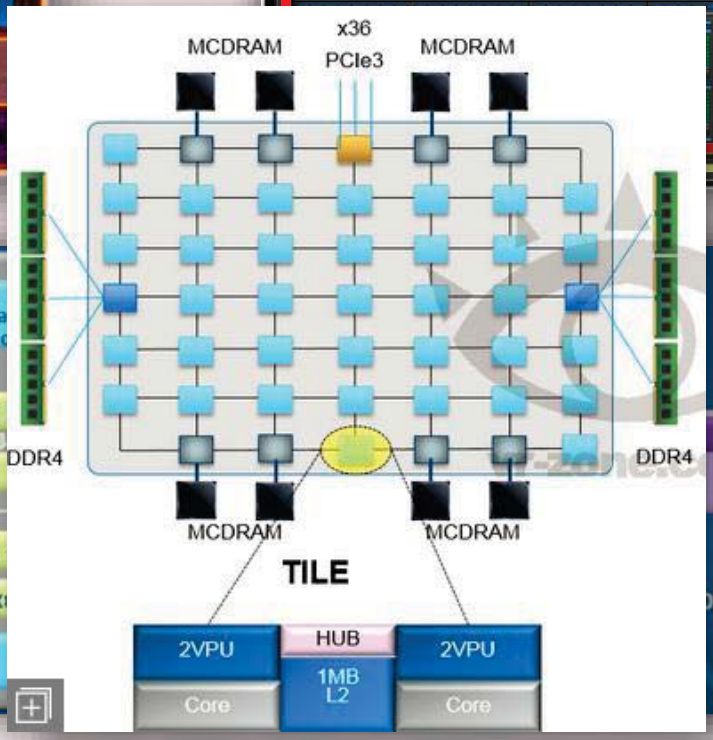
SIMD

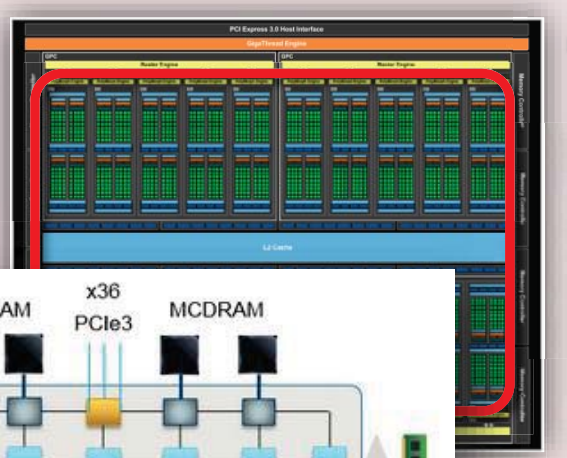
Massive

Parallel

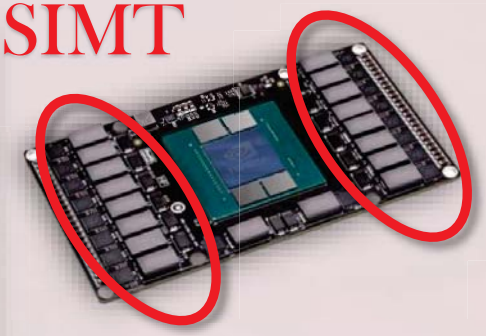
Threading



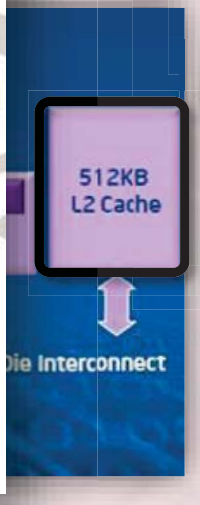


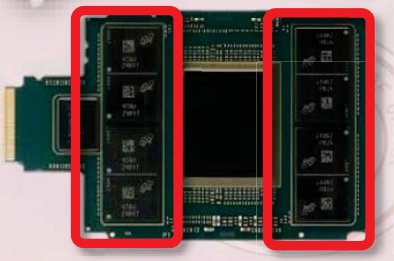


SIMT



On-board memory





TOP500和Green500中协处理器使用

TOP 10 Sites for November 2016

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
4	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
5	DOE/SC/LBNL/NERSC United States	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	622,336	14,014.7	27,880.7	3,939
6	Joint Center for Advanced High Performance Computing Japan	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Fujitsu	556,104	13,554.6	24,913.5	2,719
7	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
8	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	206,720	9,779.0	15,988.0	1,312
9	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	[illegible]	[illegible]	[illegible]	[illegible]
10	DOE/NNSA/LANL/SNL United States	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	301,056	[illegible]	[illegible]	[illegible]

Green500 List for November 2016

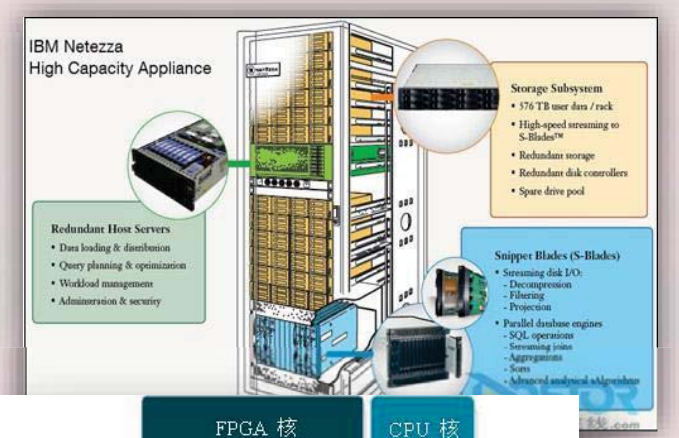
Listed below are the November 2016 The Green500's energy-efficient supercomputers ranked from 1 to 10.

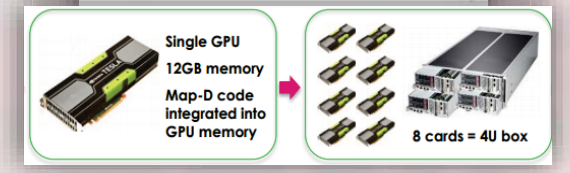
Green500 Rank	MFLOPS/W	Site	System	Total Power(kW)
1	9462.1	NVIDIA Corporation	NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100	349.5
2	7453.5	Swiss National Supercomputing Centre (CSCS)	Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100	1312
3	6673.8	Advanced Center for Computing and Communication, RIKEN	ZettaScaler-1.6, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SCnp	150.0
4	6051.3	National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	15371
5	5806.3	Fujitsu Technology Solutions GmbH	PRIMERGY CX1640 M1, Intel Xeon Phi 7210 64C 1.3GHz, Intel Omni-Path	77
6	4985.7	Joint Center for Advanced High Performance Computing	PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	2718.7
7	4688.0	DOE/SC/Argonne National Laboratory	Cray XC40, Intel Xeon Phi 7230 64C 1.3GHz, Aries interconnect	1087
8	4112.1	Stanford Research Computing Center	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	190
9	4086.8	Academic Center for Computing and [illegible] (ACCMS), Kyoto [illegible]	Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect	748.1
10	[illegible]	Thomas Jefferson National Accelerator [illegible]	KOI Cluster, Intel Xeon Phi 7230 64C 1.3GHz, Intel Omni-Path	111

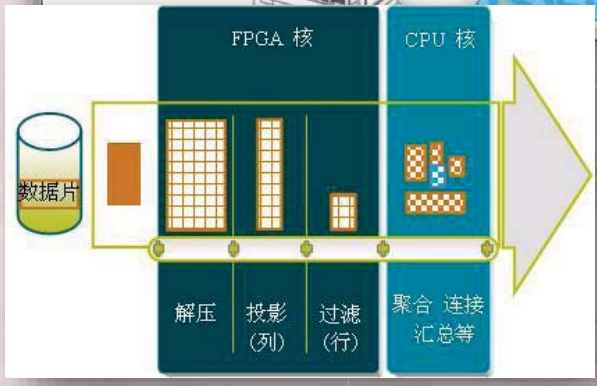
协处理器是未来高性能计算平台的主流趋势

协处理器的计算与传统CPU计算的差异

- 系统配置方案不同
  - 协处理器相对于通用CPU除了在计算性能方面有较大的优势之外，还在价格上占较大的优势。Phi和GPU相对同时期的通用CPU价格仅为50%-70%，高性能服务器通常配置少量的通用CPU和大量的协处理器提高服务器整体的性价比。
  - 未来高性能计算平台上，计算可能会全部或大部分转移到协处理器上完成，实现由CPU内存计算向协处理器内存计算的技术升级。
  - 新兴的计算型数据库甚至完全依赖协处理器计算

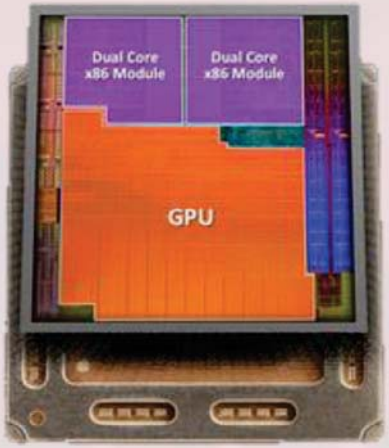


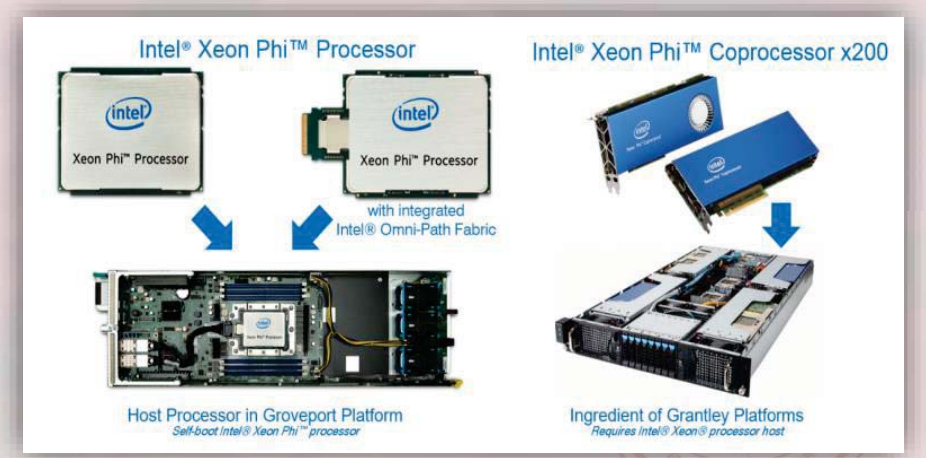




未来处理器技术发展趋势

- 从协处理器到主处理器
  - 融合技术：
    - APU：CPU与GPU融合，共享内存带宽
    - Xeon +FPGA：通过定制化加速数据处理性能
  - 统一技术：
    - Xeon Phi：成为standalone CPU，直接访问系统内存





小结

- 内存数据库技术依赖于硬件技术发展
  - 高性能处理器、大内存支持内存数据库应用
- 内存数据库需要克服硬件技术障碍
  - 内存墙：通过面向缓存的优化技术克服内存访问瓶颈
- 内存数据库通过硬件敏感技术优化性能
  - 优化多级cache访问性能：cache-conscious设计的存储模型、索引结构、数据访问及查询执行
  - 优化SIMD性能：面向SIMD的排序、连接、谓词处理等优化算法
- 新硬件技术进一步推动内存数据库发展
  - 非易失性内存NVM：简化内存数据库日志机制，优化持久性数据存储性能
  - 众核协处理器：提高内存数据库并行处理能力，分担CPU计算负载，加速计算性能
  - 新型处理器：物理融核架构支持内存数据库双查询处理引擎设计