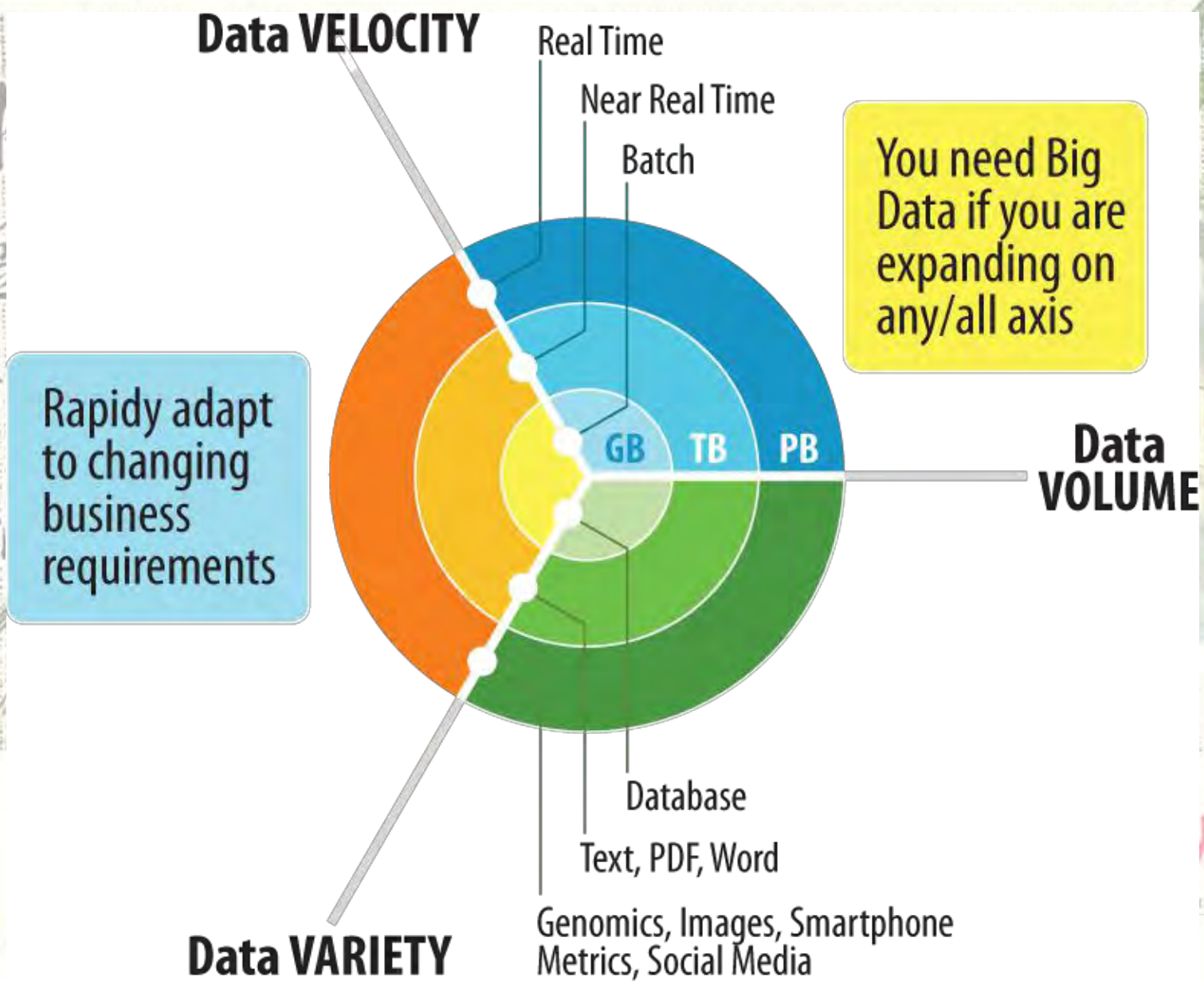
内存计算：现状、挑战与实践

金　海

华中科技大学计算机学院

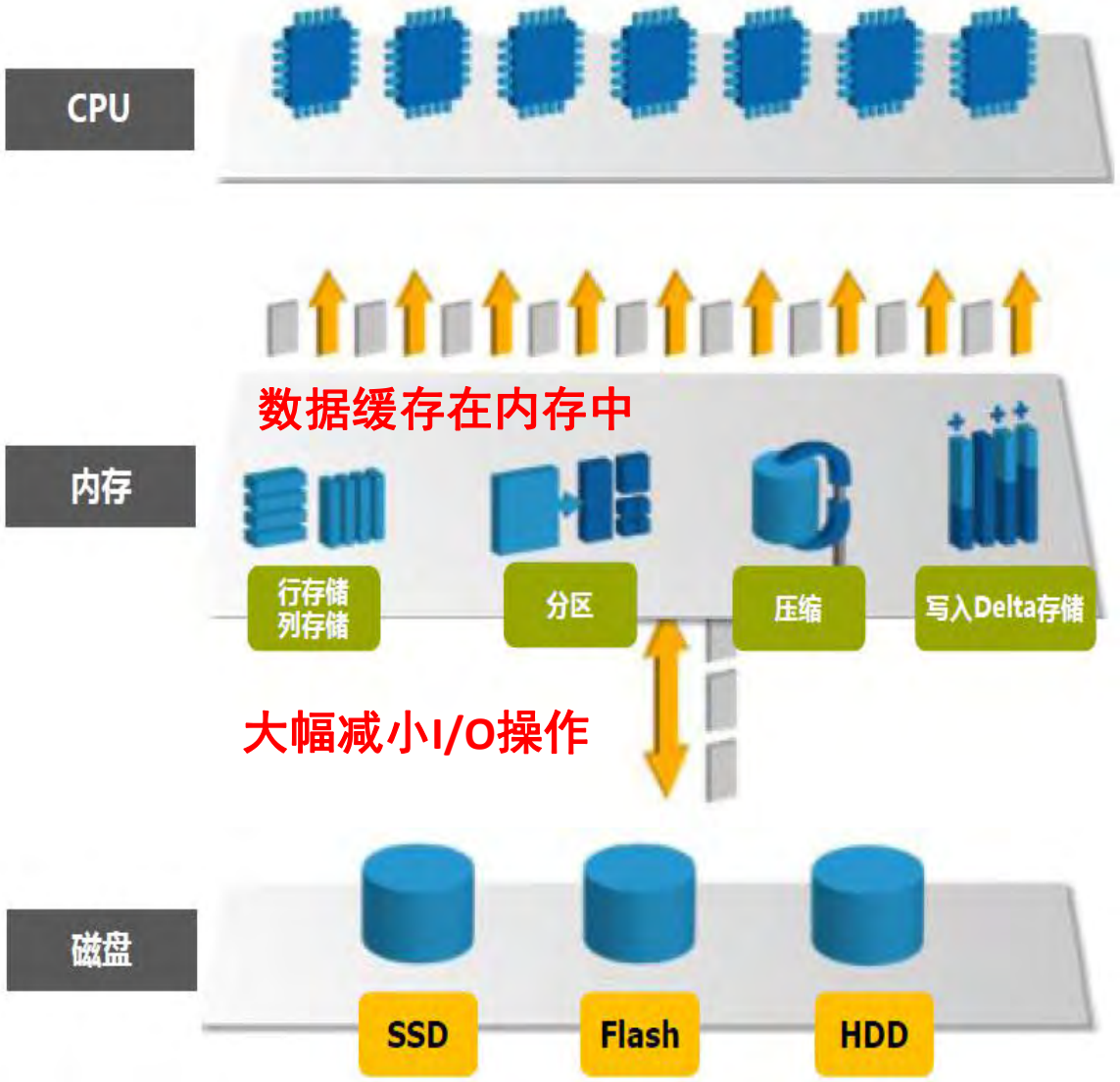
从今年的双十一看大数据处理的挑战

大数据不只是数量大

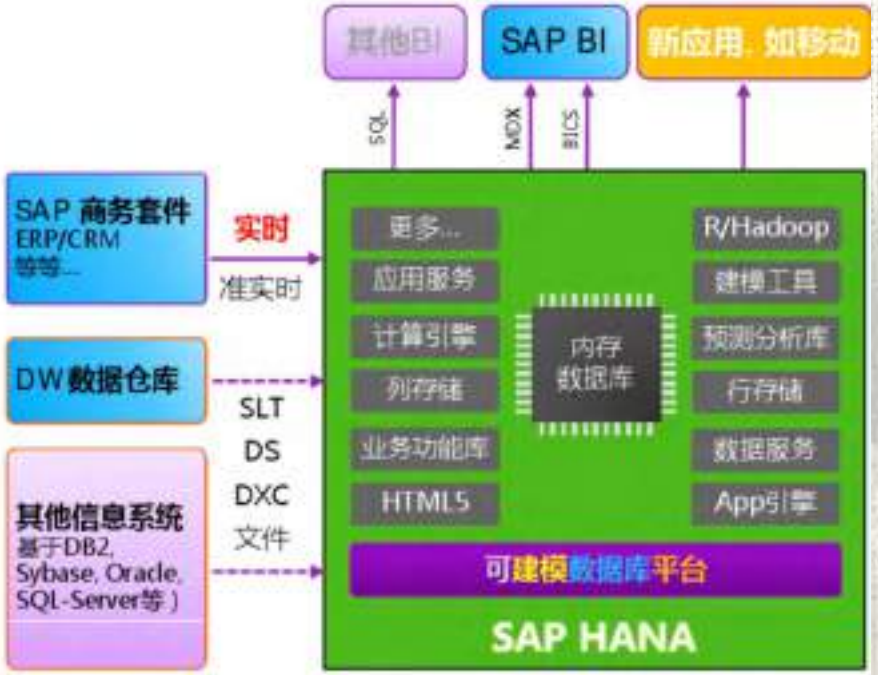

Data VELOCITY
Real Time
Near Real Time
Batch
You need Big Data if you are expanding on any/all axis
Rapidy adapt to changing business requirements
GB
TB
PB
Data VOLUME
Database
Text, PDF, Word
Genomics, Images, Smartphone Metrics, Social Media
Data VARIETY

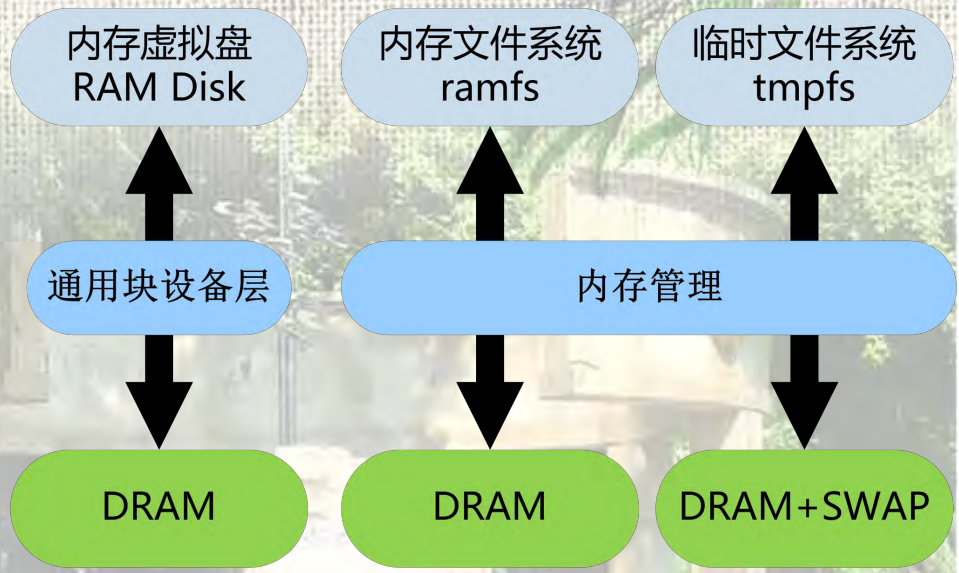
内存计算：新的计算模式


CPU
MEMORY
磁盘I/O是瓶颈


CPU
内存
数据缓存在内存中
行存储
列存储
分区
压缩
写入Delta存储
大幅减小I/O操作
磁盘
SSD
Flash
HDD

内存计算：提升数据处理时效性的重要手段





SPA HANA性能比传统磁盘系统提升408,000倍

内存文件系统I/O性能比SSD提升1000多倍

内存计算为什么现在才涌现？

- 64位CPU能够寻址16EB的数据
- DRAM的生产成本每12个月下降32%
- 市场上可提供TB级的DRAM
- 支持内存计算的软件逐渐浮现

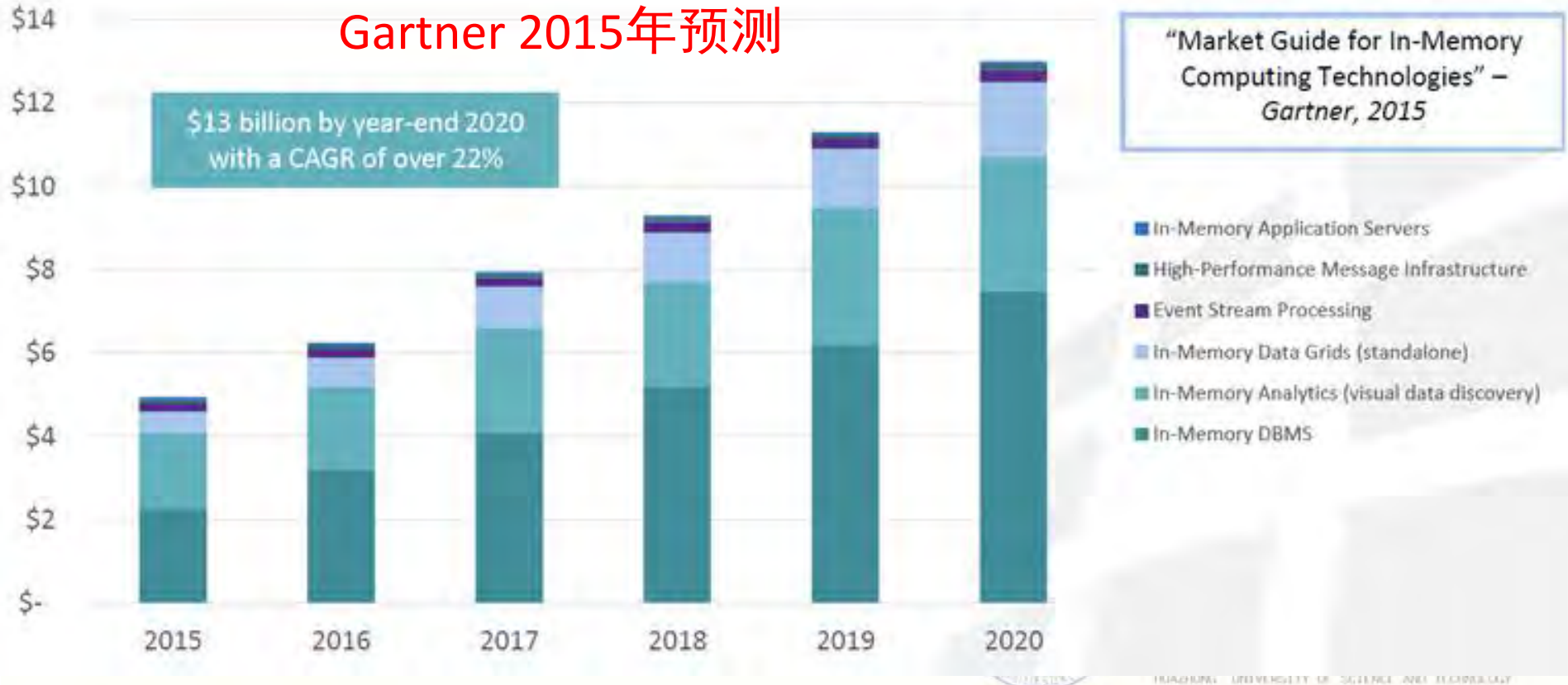
内存计算：蓬勃发展远胜预期

Gartner Says In-Memory Computing Is Racing Towards Mainstream Adoption

In-Memory Data Grid Market to Reach $1 Billion by 2016

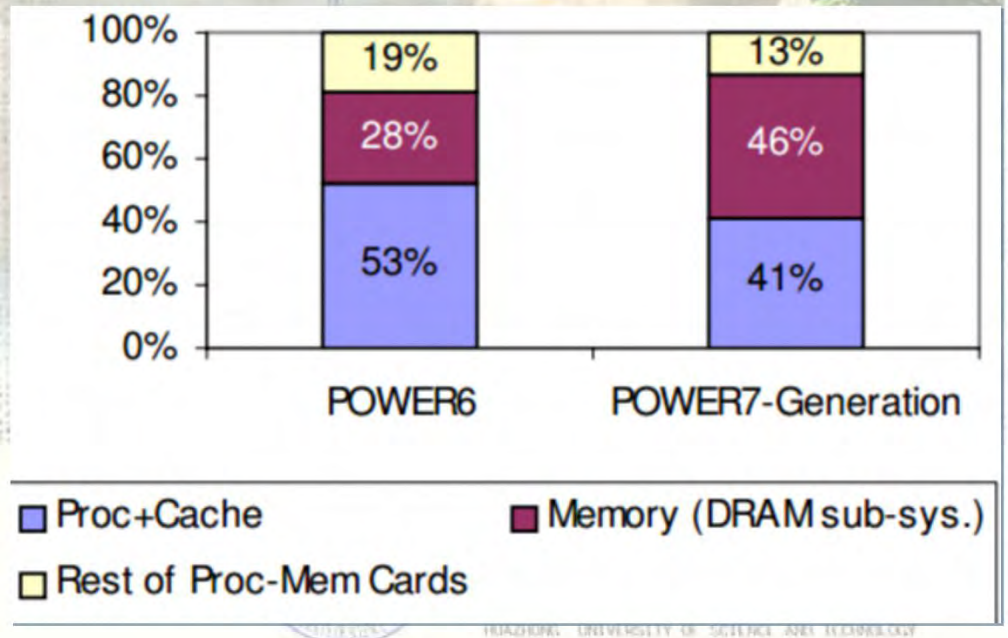
Gartner 2013年预测

Gartner 2015年预测



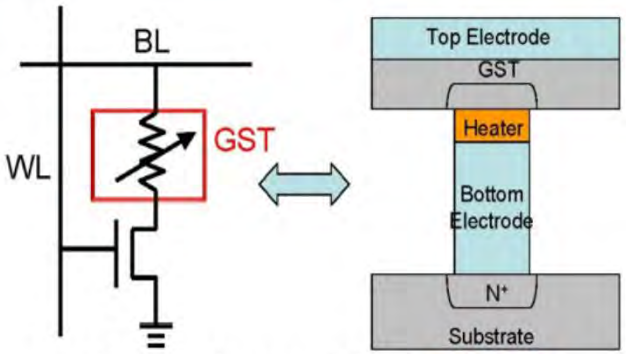
基于DRAM的内存计算模式面临的挑战

- DRAM介质易失性
  - 数据需要在内存和持久存储的外存之间交换，“数据I/O”瓶颈仍然没有完全消除
- DRAM介质存储密度低
  - 单个DRAM芯片的集成度已经接近极限，远不能满足大数据对内存容量TB级甚至PB级的需求
- DRAM功耗高
  - 随着内存容量的增大，DRAM内存功耗占系统总能耗的比例可高达46%，静态功耗（刷新操作）占DRAM总功耗的比例接近50%
- 内存子系统成本高
  - 单位存储容量的内存价格高昂

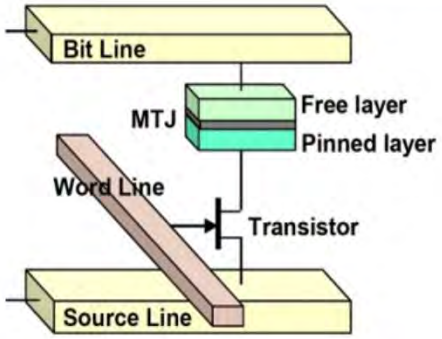


IBM POWER7子系统能耗比例

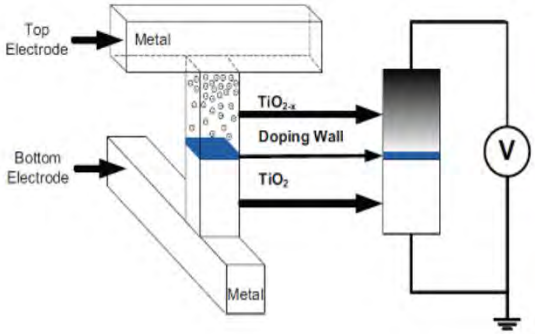
非易失存储技术的兴起

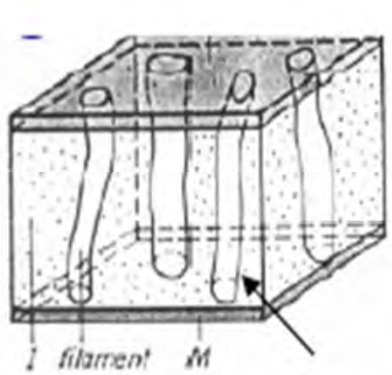


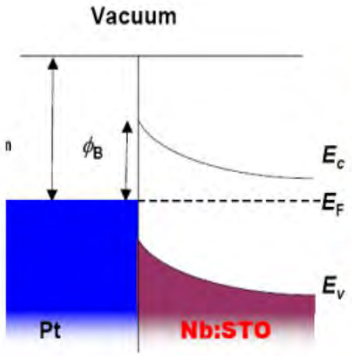
Phase-change RAM (**PCRAM**)

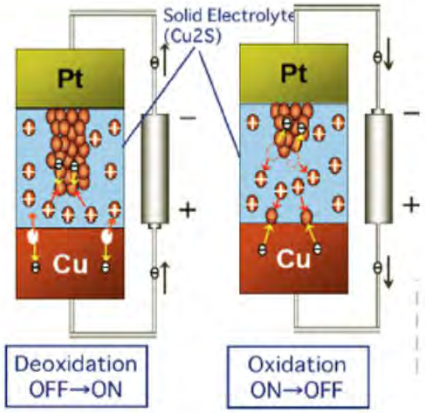


Magnetic RAM (**MRAM**)



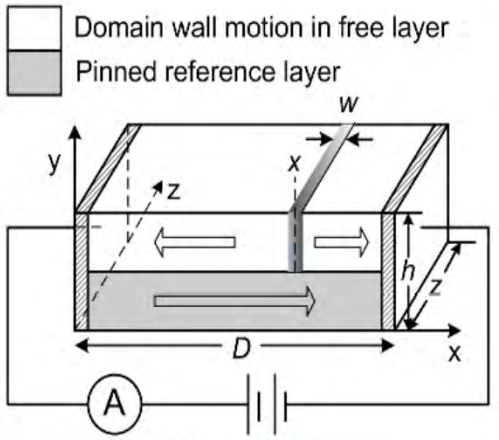






Resistive RAM (**RRAM**)

From left to right: Filament-based, Interface-based and PMC



Memristor

Top: thin-film device

Bottom: Spintronic material

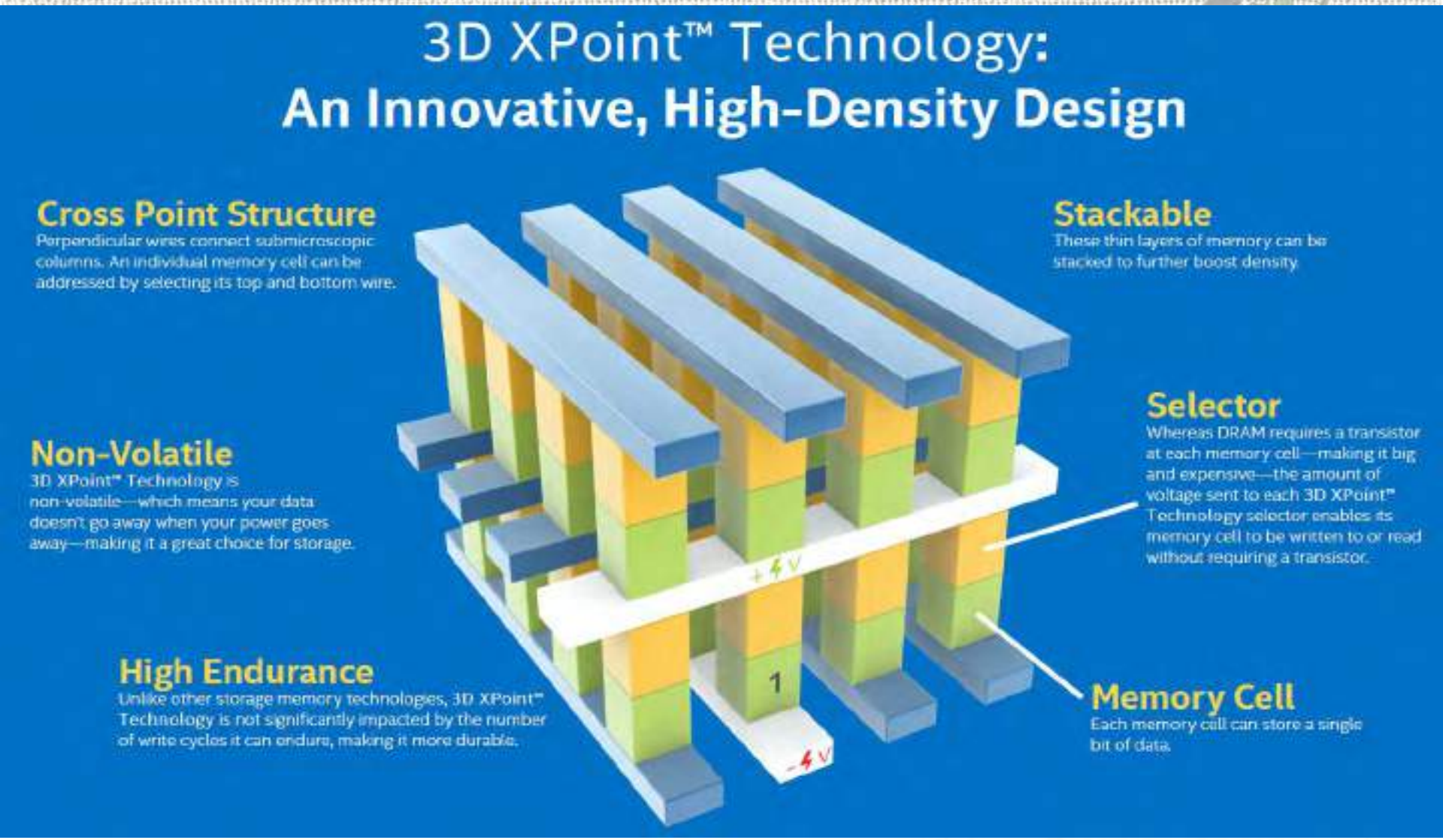
Emerging Non-volatile Memory (NVM) Technologies

新型存储介质的性能不断逼近传统DRAM

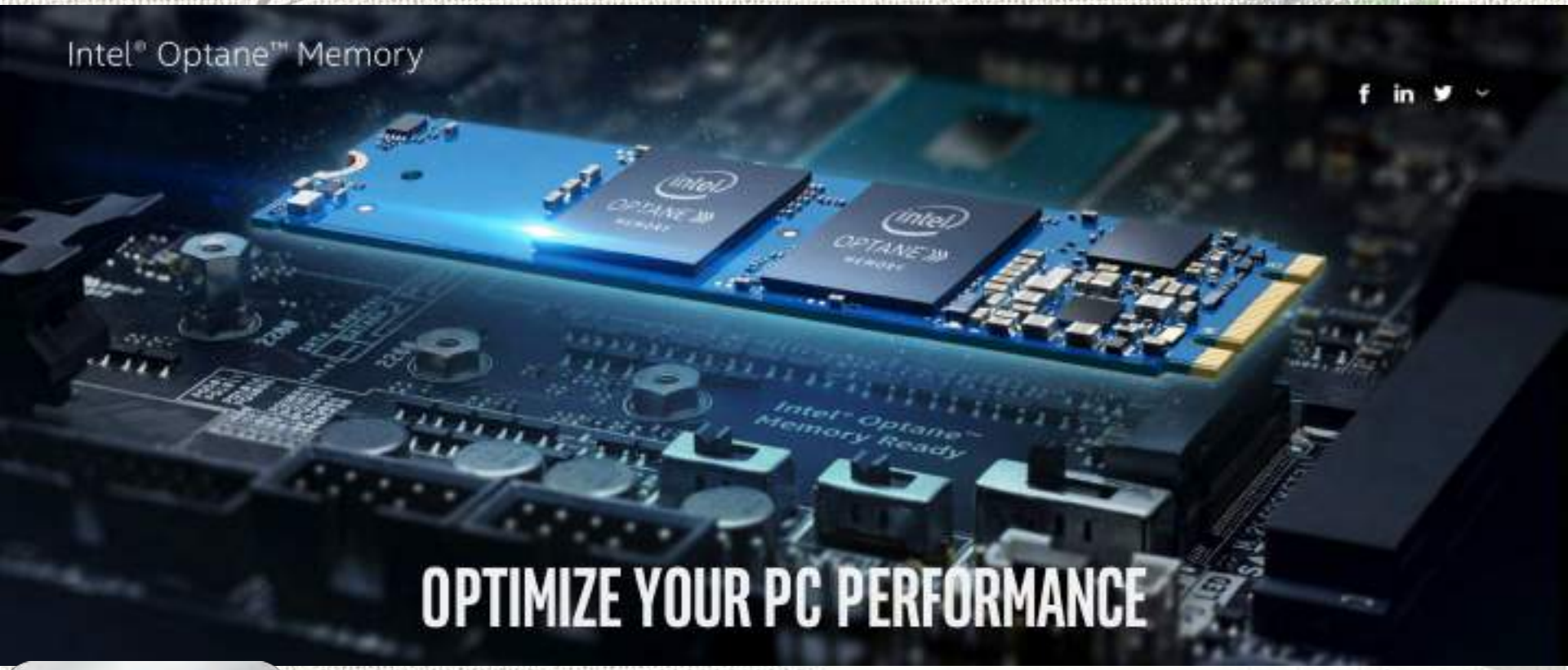
- ITRS（International Technology Roadmap for Semiconductors：国际半导体技术蓝图）Projection

	Memristor	PCM	STT-RAM	DRAM	Flash	HD
Chip area per bit (F^2)	4	8–16	14–64	6–8	4–8	n/a
Energy per bit (pJ)[2]	0.1–3	2–100	0.1–1	2–4	10^1–10^4	10^6–10^7
Read time (ns)	<10	20–70	10–30	10–50	25,000	5–8x10^6
Write time (ns)	20–30	50–500	13–95	10–50	200,000	5–8x10^6
Retention	>10 years	<10 years	Weeks	<Second	~10 years	~10 years
Endurance (cycles)	~10^{12}	10^7–10^8	10^{15}	>10^{17}	10^3–10^5	10^{15} ?
3D capability	Yes	No	No	No	Yes	n/a

Intel-Micron 3D XPoint Technology

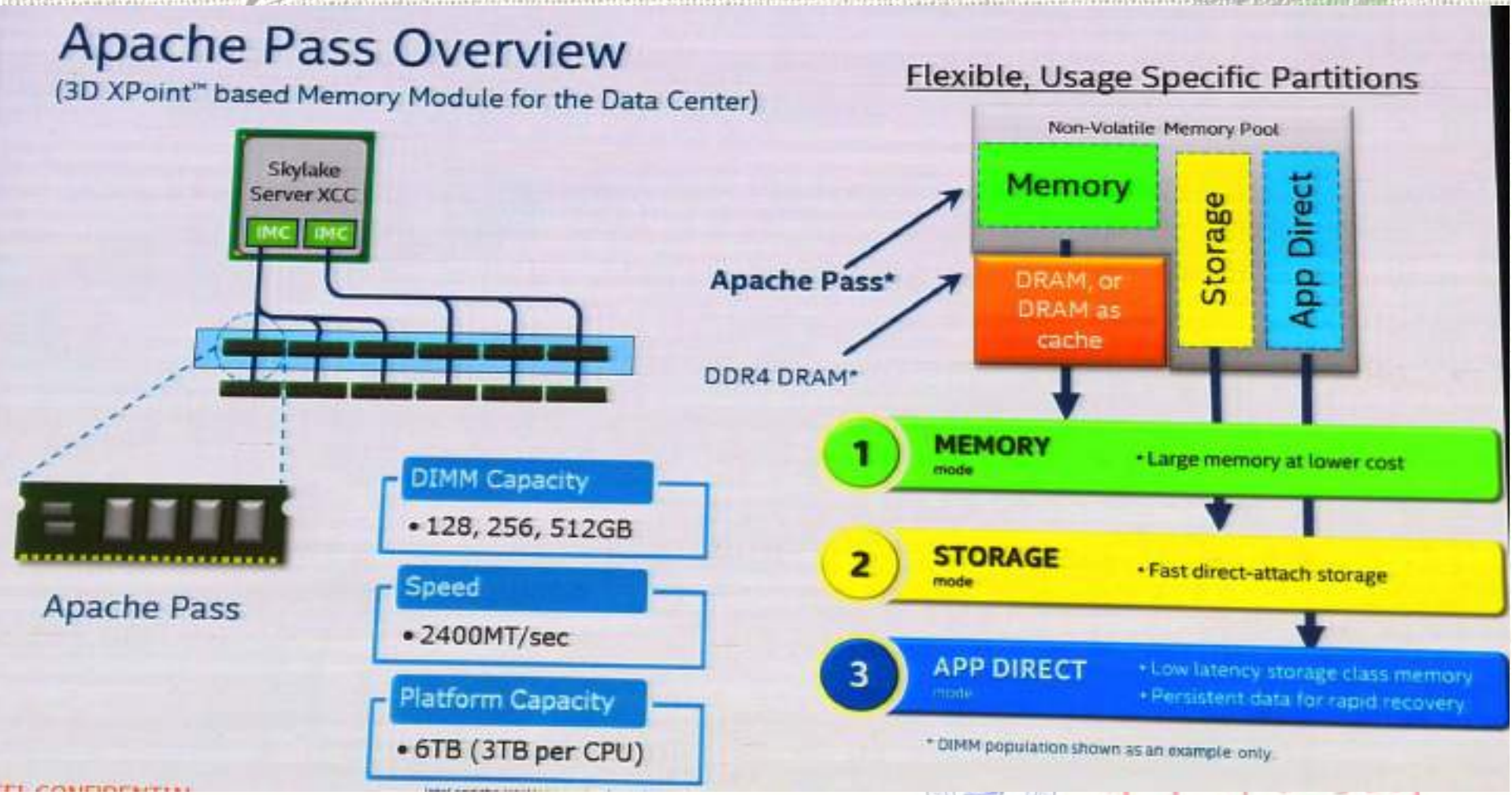


Intel Non-Volatile Memory Solutions



華中科技大學

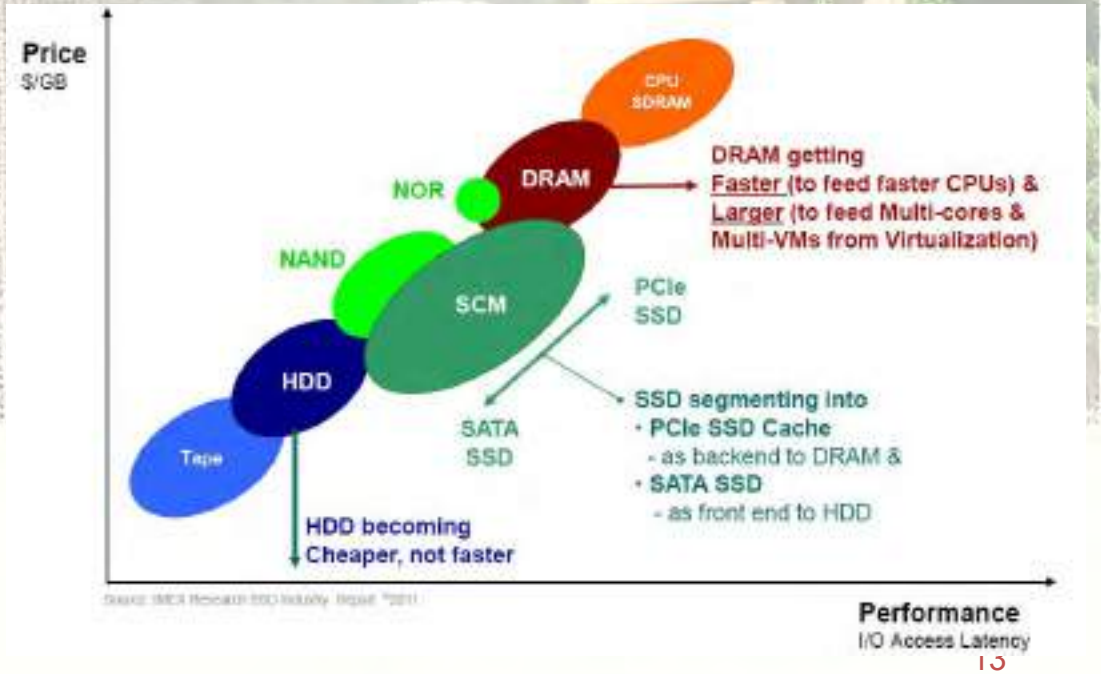
Intel Non-Volatile Memory Solutions



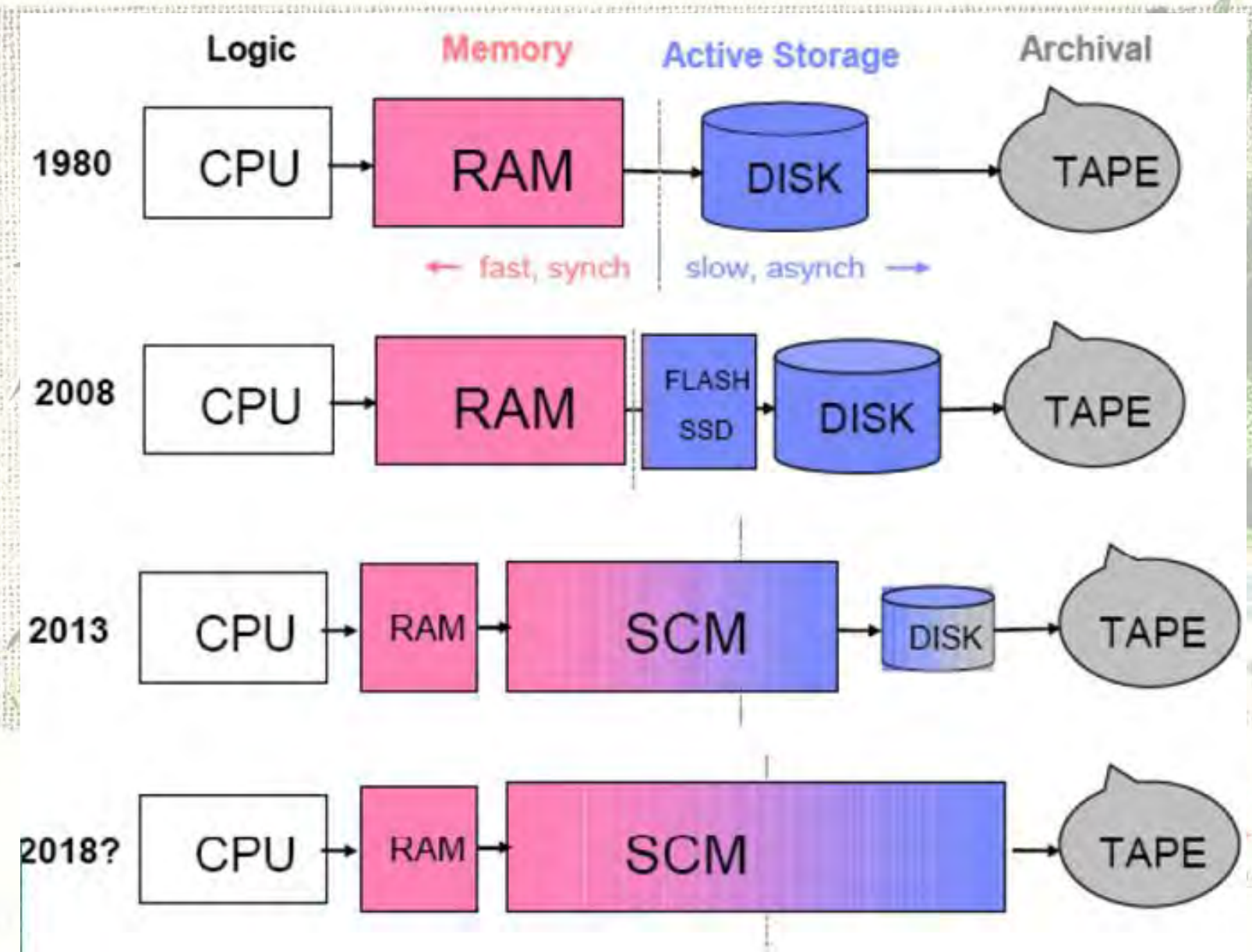
新型存储级内存（SCM）为内存计算带来曙光

新型存储级内存（Storage Class Memory）关键特性

- 可按字节寻址（可作内存用）
- 持久存储，断电数据不丢失
- 比NAND flash读写快1000倍，读延迟接近DRAM，写延迟高1倍
- 耐久性比NAND flash高1000倍，但存在写寿命上限
- 存储密度比NAND flash高1000倍，比DRAM高一个数量级
- 静态功耗接近“零”

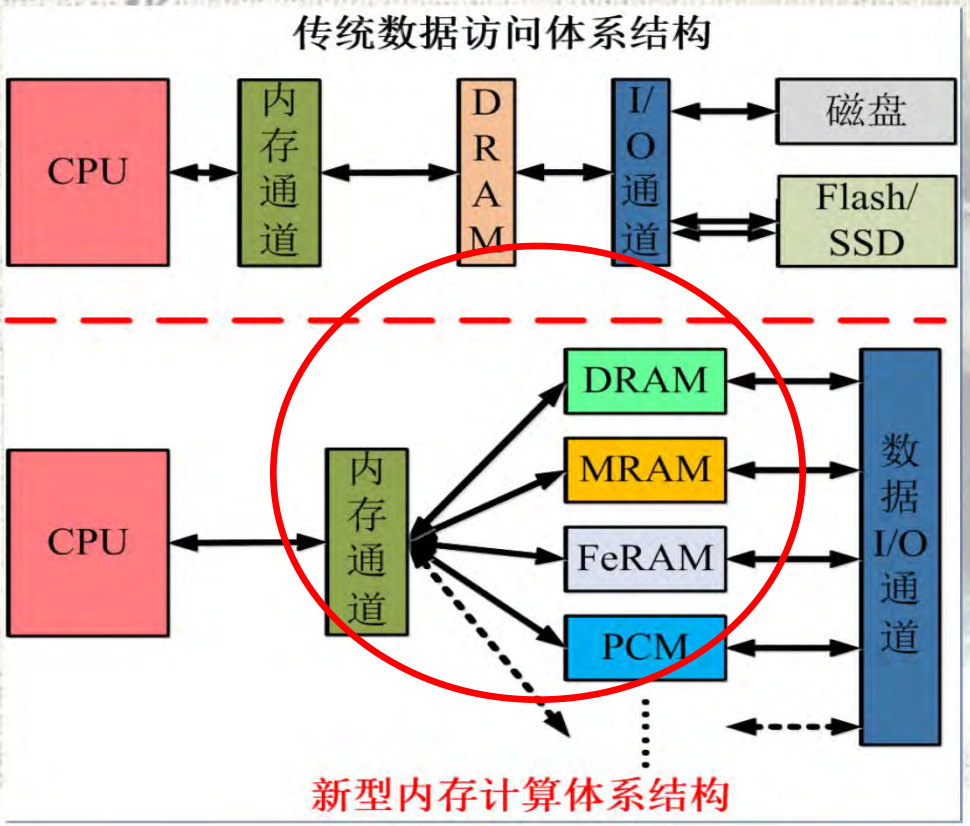


重构虚拟内存结构：打破I/O瓶颈



内存计算：新结构

构建DRAM+SCM的混合层次/并行内存结构



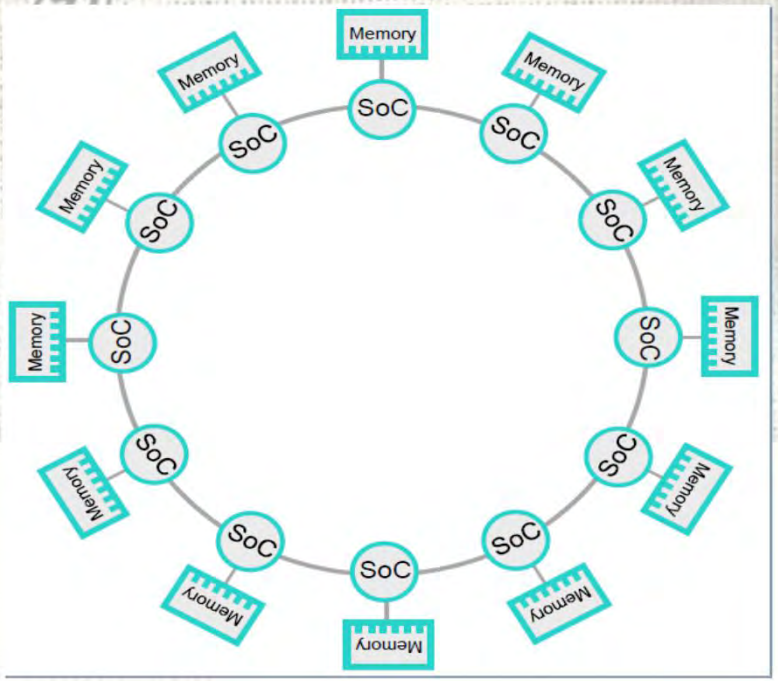
- “内存”：混合异构内存体系
- 多通道“内存”访问能力
- 计算不仅在DRAM上，也在SCM等新“内存”部件上
- 数据不仅在DRAM上，也在SCM等新“内存”部件上

从以计算为中心转向计算与数据相结合的全新内存计算架构

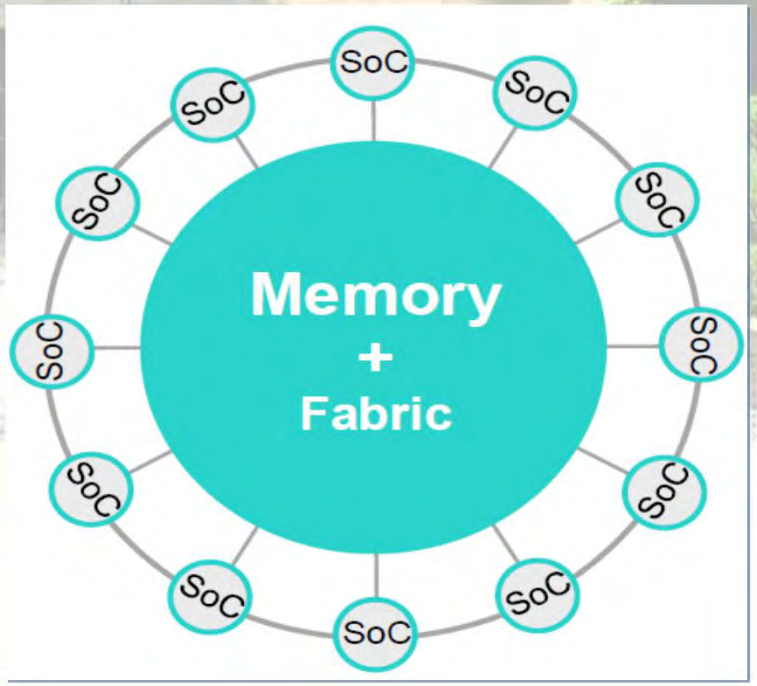
HP: The Machine

- 采用了非易失性存储技术—忆阻器
- 单机柜40个节点可共同分享160TB的内存资源，是当前世界上最大的单机内存计算系统

CPU为中心的计算



内存为中心的计算

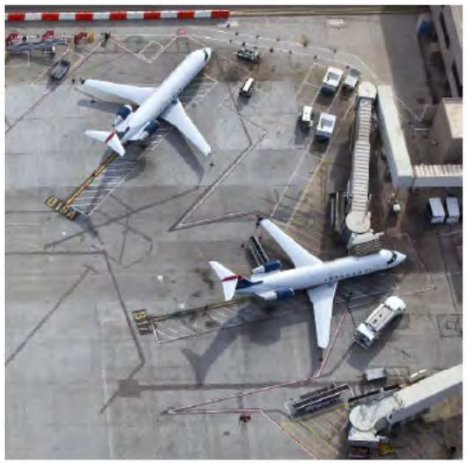


HP: The Machine

Modify existing frameworks

New algorithms

Completely rethink



In-memory analytics	Similarity search	Large-scale graph inference	Financial models
15x faster	**20x** faster	**100x** faster	**8,000x** faster

源自： HP高级工程师Dr. Kimberly Keeton在MSST'17上的特邀报告

SNIA NVM Programming TWG

- Members
  - EMC, Fujitsu, Fusion-io, HP, HGST, Inphi, Intel, Intuitive Cognition Consulting, LSI, Microsoft, NetApp, PMC-Sierra, Qlogic, Red Hat, Samsung, Seagate, Sony, Symantec, Viking, Virident, VMware
  - Calypso Systems, Cisco, Contour Asset Management, Dell, FalconStor, Hitachi, Huawei, IBM, IDT, Marvell, Micron, NEC, OCZ, Oracle, SanDisk, Tata Consultancy Services, Toshiba

SNIA NVM Programming Model v1.2

SNIA

Advancing storage & information technology

NVM Programming Model (NPM)

Version 1.2

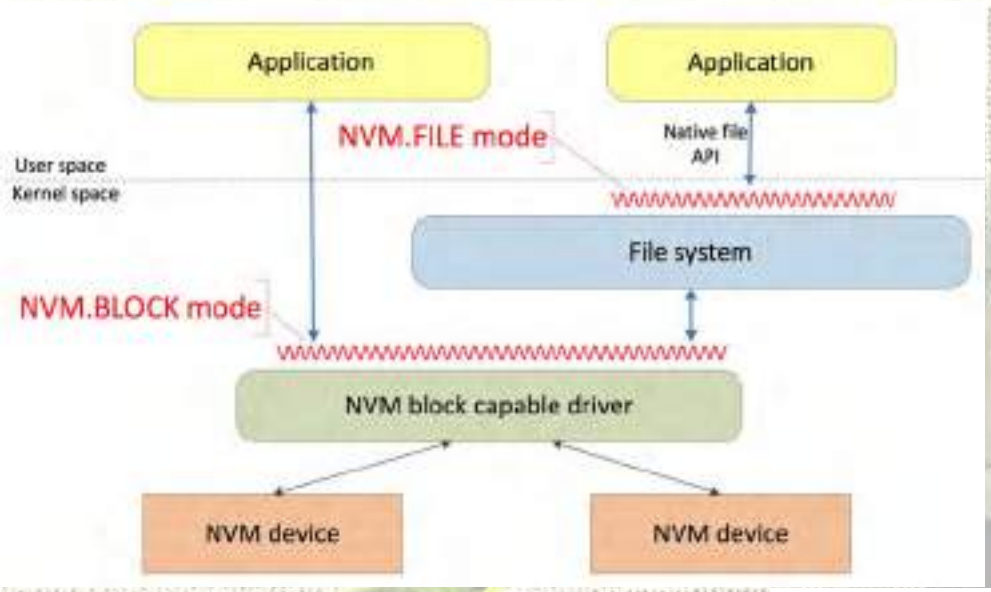
Abstract: This SNIA document defines recommended behavior for software supporting Non-Volatile Memory (NVM).

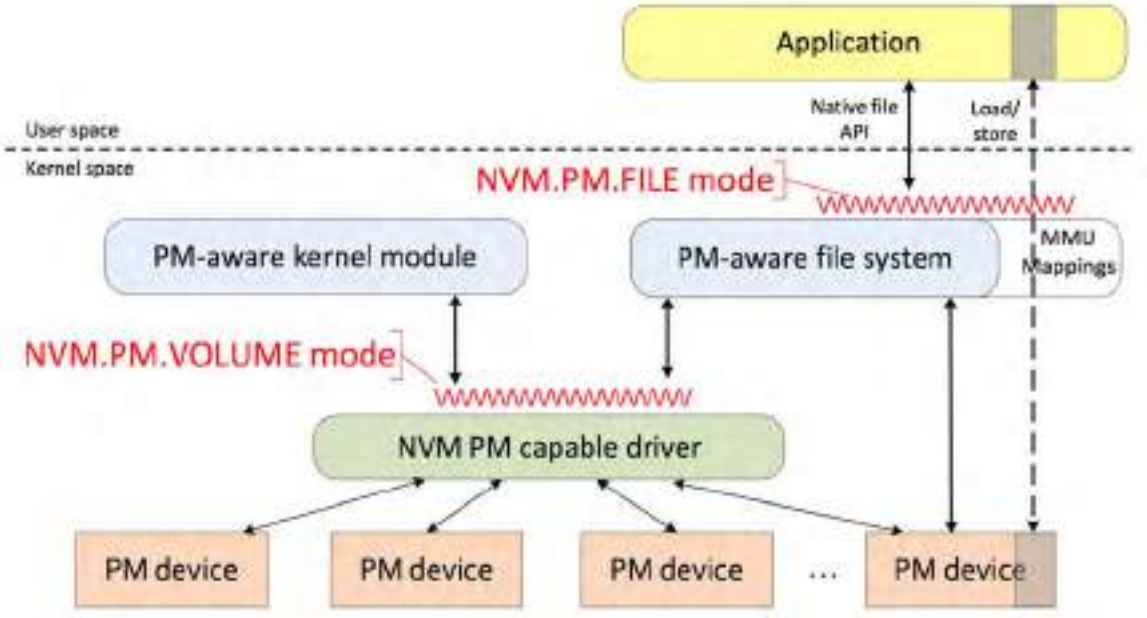
This document has been released and approved by the SNIA. The SNIA believes that the ideas, methodologies and technologies described in this document accurately represent the SNIA goals and are appropriate for widespread distribution. Suggestion for revision should be directed to http://www.snia.org/feedback/.

SNIA Technical Position

June 19, 2017

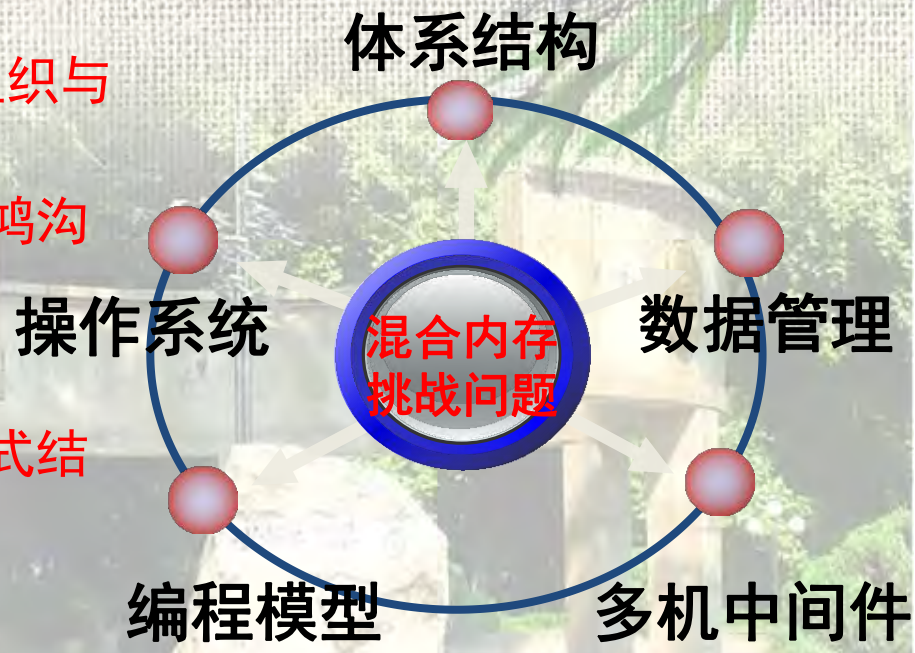
Two NVM Programming Modes Overview





混合内存架构给软硬件生态系统带来巨大冲击

- **体系结构**
  - 异构内存部件（NVM/DRAM）的组织与管理（多通道、多层次？）
  - 多核与大内存之间不断加大的带宽鸿沟（内存级并行?）
- **操作系统**
  - 大容量内存的寻址问题（大页？段式结构？）
  - 面向混合内存的操作系统任务调度
- **数据组织**
  - 文件结构、key/value store、数据库？
- **编程模型**
  - 无需考虑磁盘I/O，数据持久化，计算存储融合架构下的编程模型



- **多机中间件**
  - 内存池化，异构内存网络

非易失存储技术的安全挑战

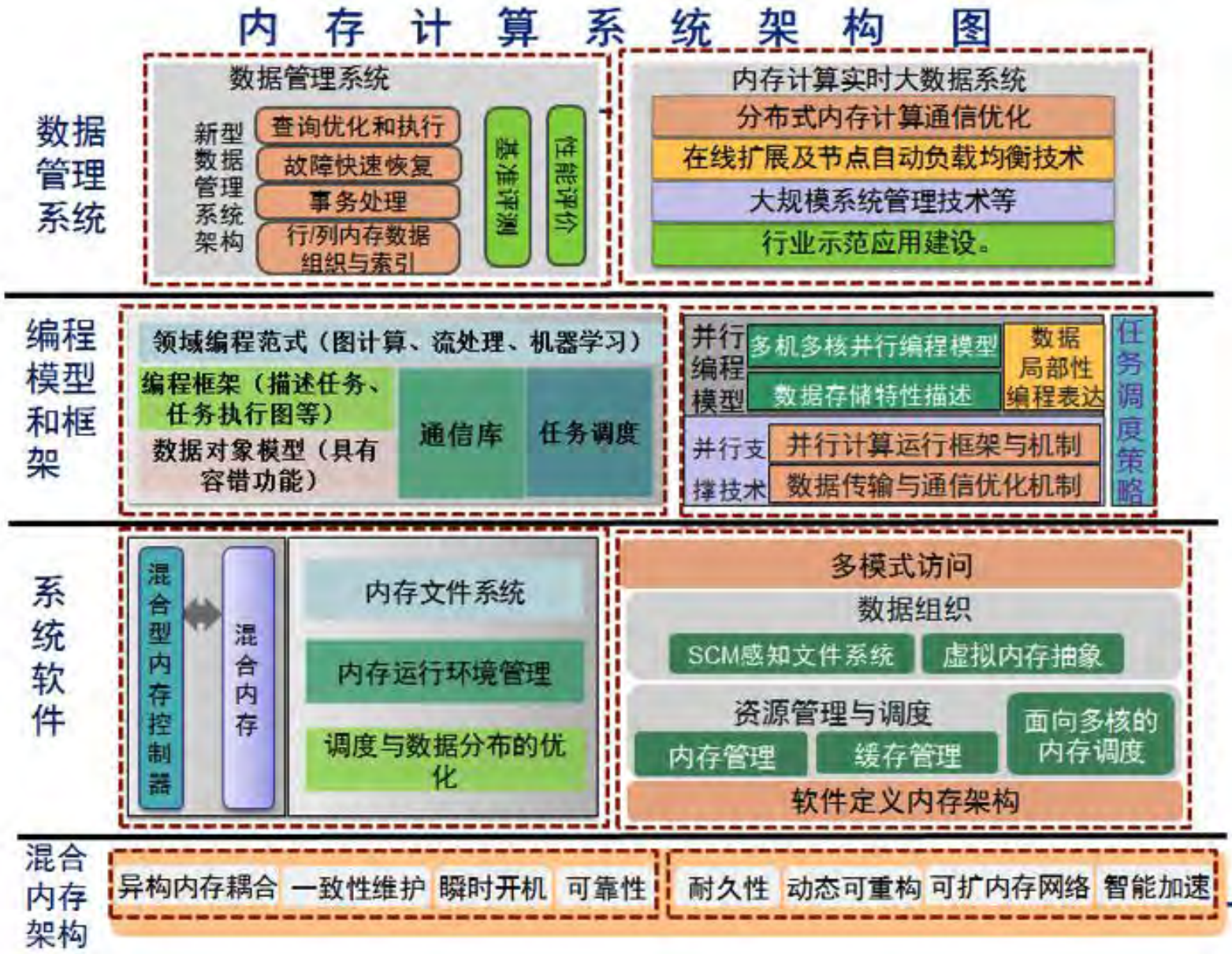
- Limited endurance → Wearout attacks
  - Better architecting of memory chips to absorb writes
  - Hybrid memory system management
  - Online wearout attack detection
- Non-volatility → Data persists in memory after powerdown
  → Easy retrieval of privileged or private information
  - Efficient encryption/decryption of whole main memory
  - Hybrid memory system management
- Multiple bits per cell → Information leakage (via side channel)
  - System design to hide side channel information

面向大数据的内存计算关键技术与系统

- 异构混合内存体系结构研究与开发
- 内存计算系统软件研究与开发
- 基于内存计算的并行处理系统研究与开发
- 基于内存计算的数据管理系统研究与开发

华中科技大学
HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

面向大数据的内存计算关键技术与系统



混合内存计算系统软件


需 求
可靠的内存管理
一致的数据组织
能耗与性能权衡
多样化应用需求
设 计
统一内存管理
持久区域管理
感知应用数据
感知内存介质
缓存管理
多核内存调度
多模式访问
软件定义系统架构
模拟器扩展
系统软件
动态自适应混合内存管理系统
SCM感知文件系统
新型异构内存Cache替换策略
多核环境的异构内存调度优化
多模式访问接口架构
五项关键技术
可重构的系统架构
系统架构
内存计算全系统模拟平台
模拟平台
异构混合内存计算系统原型
原型系统