高速、智能、开放：中国开源文件系统（COFS）助力下一代存储系统的发展

曾令仿

新型智能计算系统研究中心

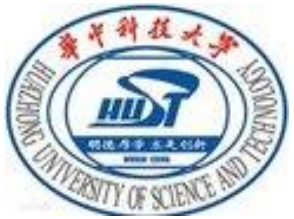
之江实验室

2020年10月

COFS
China Open File System

个人介绍

- 主要学习和工作经历
  - 2020/05–至今　之江实验室，新型智能计算系统研究中心，研究员
  - 2009/02–2020/04　华中科技大学，武汉光电国家研究中心，副教授
  - 2009/02–2020/04　华中科技大学，计算机科学与技术学院，副教授
  - 2016/05–2018/05　德国美因茨大学，数据中心，Visiting Professor
  - 2010/11–2013/07　新加坡国立大学，电子工程系，Research Fellow
  - 2007/06–2008/08　新加坡国立大学，电子工程系，Research Fellow
  - 2006/05–2009/01　华中科技大学，计算机科学与技术学院，讲师
  - 2003/09–2006/06　华中科技大学，计算机系统结构，博士研究生
- 主要研究方向
  - 高性能存储系统（芯片级、器件级、设备级、系统级等软硬件）
  - 大数据与计算智能（AI for Storage、Storage for AI等）
  - 隐私增强信息存储（数据安全删除、ORAM、区块链存储等）



汇报提纲

- 中国开源文件系统（COFS）介绍
- COFS委员会创始成员介绍
- 浅谈下一代存储系统

COFS
China Open File System

中国开源文件系统（COFS）介绍

- COFS宗旨、职责、权力、决策机制
  - 服务于中国：开源文件系统相关（公益）活动（例如：开发、维护、研讨、用户交流等）
  - 成员通过不定期会议投票决策COFS相关事务
  - 主办CLUG（China Lustre User Group meeting）
  - BeeGFS、Ceph、Gluster等开源文件系统
  - 官方网站：http://www.chinafs.org/

COFS
China Open File System

汇报提纲

- 中国开源文件系统（COFS）介绍
- COFS委员会创始成员介绍
- 浅谈下一代存储系统

COFS委员会创始成员介绍

- COFS五位创始成员（以汉语拼音排序）
  - **董勇**，国防科技大学计算机学院计算机研究所副研究员
  - **李希**，DDN/Whamcloud公司首席工程师
  - **汪璐**，中国科学院高能物理研究所副研究员
  - **薛巍**，清华大学计算机系副教授，地球系统科学系双聘副教授，高性能计算研究所所长
  - **曾令仿**，之江实验室新型智能计算系统研究中心研究员

COFS委员会创始成员：董勇

董勇，男，博士，国防科技大学计算机学院副研究员，硕士生导师。长期从事国产高性能计算机系统研制。主要研究方向为大规模并行计算环境，包括并行文件系统、高性能互连通信、资源管理系统等。获省部级科技进步特等奖1项，一等奖2项。在IEEE TC，SC等国内外学术期刊、会议发表学术论文20余篇，授权专利近20项。参与国家自然科学基金、863、国家重点研发计划、军队预研等方面重点项目多项。

COFS
China Open File System

COFS委员会创始成员：李希

李希，DDN/Whamcloud首席工程师，亚太区研发主管。为Lustre文件系统设计并实现了策略引擎、客户端持久缓存、Project Quota、QoS等多种功能，为Lustre、Ext4、MpiFileUtils、LustrePerfMon等开源项目贡献了大量代码。

COFS
China Open File System

COFS委员会创始成员：汪璐

汪璐，中国科学院高能物理研究所副研究员,主要研究方向为海量存储、智能运维等，在本单位负责面向高能物理数据密集型计算的分布式文件系统架构、建设和运维，具有10余年的Lustre社区版运维经验。

COFS
China Open File System

COFS委员会创始成员：薛巍

薛巍，清华大学计算机系副教授，地球系统科学系双聘副教授；清华计算机系高性能计算研究所所长。主要研究兴趣包括大规模科学计算、不确定性量化分析。曾获ACM Gordon Bell Prize、清华大学-浪潮集团计算地球科学青年人才奖、教育部科学技术进步一等奖、中国电子学会电子信息科学技术奖一等奖。

COFS委员会创始成员：曾令仿

曾令仿，之江实验室研究员。 2006年博士毕业于华中科技大学并留校任教直至2020年5月加入之江实验室。曾在新加坡国立大学和德国美因茨大学工作六年。2006年获世界超级计算机大会高性能存储挑战竞赛决赛奖（排名第一）。2011年获湖北省技术发明一等奖（排名第四）。CCF高级会员与CCF2019杰出演讲者。

COFS
China Open File System

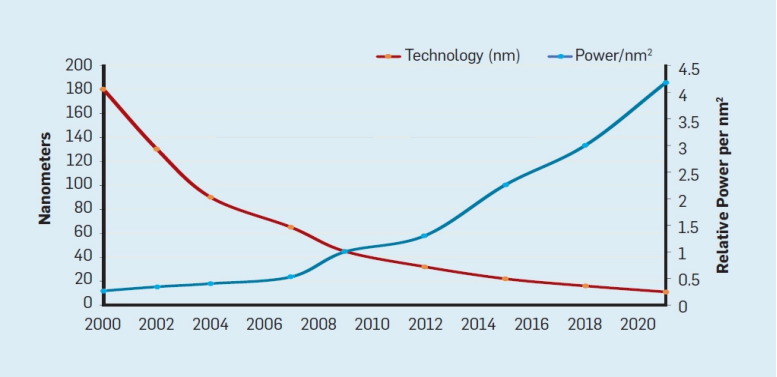
汇报提纲

- 中国开源文件系统（COFS）介绍
- COFS委员会创始成员介绍
- 浅谈下一代存储系统

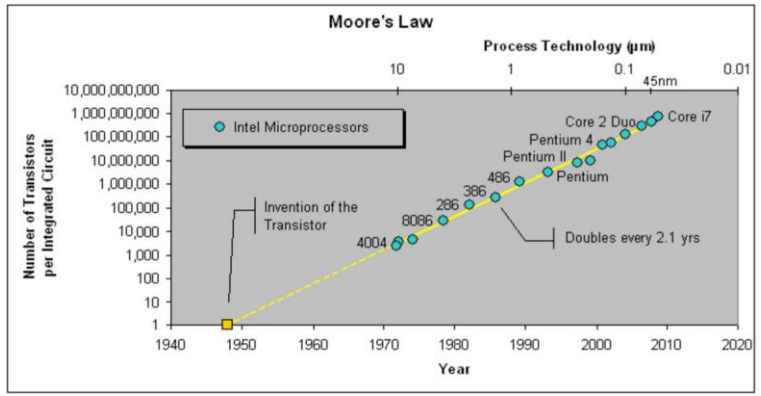
浅谈下一代存储系统

- 计算体系结构领域“定律”

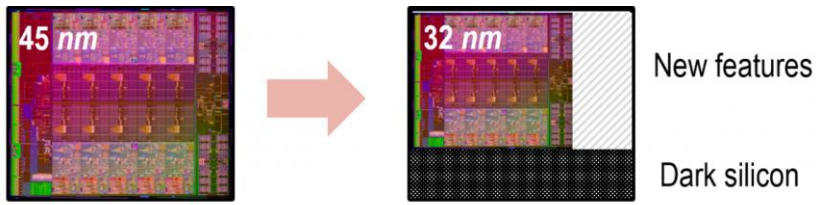
登纳德Dennard缩放定律



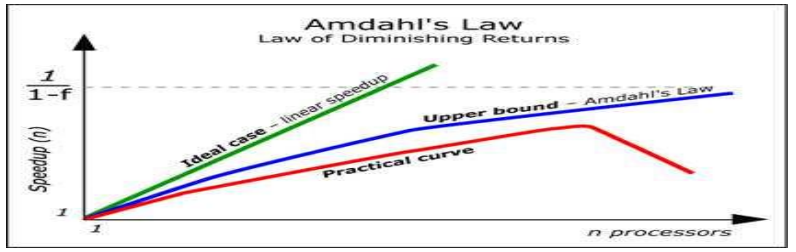
摩尔定律Moore's law



暗硅现象Dark silicon

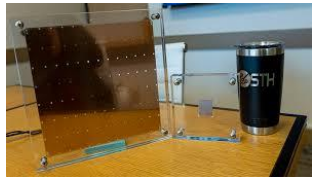


阿姆达尔Amdahl定律



浅谈下一代存储系统

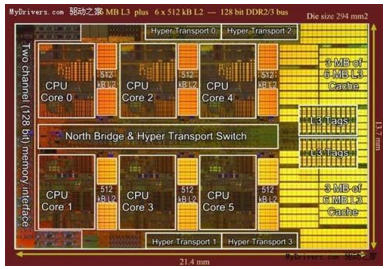
- 计算、网络及存储领域正发生全方面的剧烈变化



处理器（多核化、异构化、晶圆级）

X86性能增长变缓，CPU/GPU/FPGA异构计算盛行

从独立的异构计算向异构融合演进



网络（极低延时化、大规模化）

全SSD时代促进RoCE协议、网络普及，实现NVMe数据存储

进入百纳秒级极低时延网络，让算力更贴近数据

CCIX SEAMLESS ACCELERATION

CXL Compute Express Link

OpenCAP

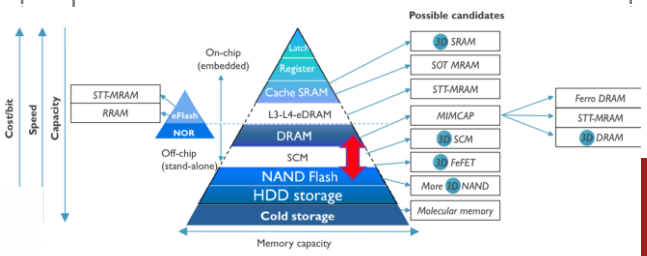
GEN Z

介质（两头延伸、每比特极致性价比）

工艺达极限，进入技术瓶颈期，成本高，内存墙效应明显

3DXPoint、MRAM、PCM

QLC/PLC/OLC、DNA



COFS China Open File System

浅谈下一代存储系统

- 信息存储面临重大挑战
  - 海量数据：数据量爆炸式剧增，各类应用层出不穷
  - 能耗瓶颈：数据存放及访问带来巨大的能耗
  - 处理挑战：Big Data、Big Storage、AI
  - 安全可信：存储介质、存储协议、存储网络等环境复杂
  - 长期保存：可用性、可靠性、性能、安全、成本等
  - 服务质量：低延迟、高通量、低成本、高可靠等

“长期存在，更加突出”

COFS
China Open File System

浅谈下一代存储系统

- 信息存储面临重大机遇
  - 新型非易失内存：3DX Point、MRAM、RRAM
  - 3D封装：3D NAND Flash、3D Stack DRAM
  - 存算一体：Flash Memeory、MRAM、RRAM，生物计算
  - 智能存储：AI for Storage、Storage for AI
  - 高速互联：NVLink&NVSwitch，NVMe、RDMA，5G
  - DNA存储：端到端（廉价）全自动系统已实现突破
  - 玻璃存储：从5维扩展到更高维

浅谈下一代存储系统

●索引结构是信息存储领域关键技术之一

索引分类	面向NVRAM的索引结构
树形索引	NV-Tree[23]，wB+-Tree[24]，FPTree[25]，WORT[26]，FAST&FAIR[27]，NoveLSM[28]，clfB-tree[29]，Circ-Tree[30]，DPTree[31]，RNTree[32]，LB+-Trees[33]，Adapting B+-Tree[34]，Tree-Based Address Mapping[35]，NV-Skiplist[36]
散列索引	FPHT[37]，Path Hashing[38]，Level Hashing[21]，Group Hashing[39]，CCEH[40]，Dash[41]
混合索引	HiKV[42]，HART[43]，HMEH[49]
索引持久化范式与框架	RECIPE[44]，NVTraverse[45]，PMwCAS[46]，Pronto[47]，MOD[48]

参考：程稳 et al.，新型存储技术与索引结构研究进展，技术调研报告，华中科技大学，2020.09

浅谈下一代存储系统

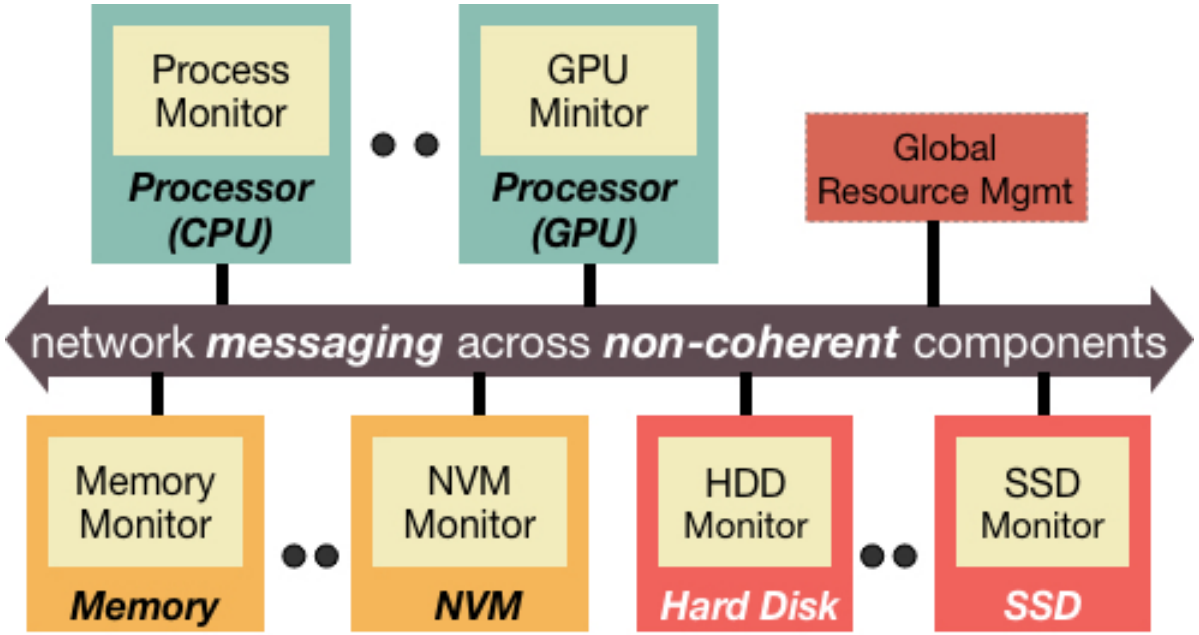
- 文件系统是信息存储领域关键技术之一
  - 节点级
    - ✓磁带、磁盘、闪存、非易失内存、主存、DNA等文件系统
    - ✓手机、平板、个人电脑等终端
    - ✓服务器
  - 分布式
    - ✓集群文件系统
    - ✓广域网文件系统

下面从文件系统角度浅谈存储系统发展趋势

COFS
China Open File System

下一代存储系统趋势一：适应资源分解

- **“文件系统”中文件系统**
  - 计算单元、内存、存储扁平化，扩展灵活
  - 高速网络互联
  - 计算容器化、无状态化
  - 内存虚拟化
  - 存储虚拟化
  - 统一资源管理与调度
  - 故障处理透明
  - 应用透明



参考： https://cseweb.ucsd.edu/~yiying/disaggregate.html

趋势二：面向特定领域/应用/功能

- **Log/Read-Only/Long-Term/……文件系统**
  - 安全
  - 数据一致性
  - 高性能
  - 高可靠/高可用
  - 高扩展
  - ……

趋势三：“智能”文件系统

- **“统一”文件系统**
  - 不同存储介质（光盘、磁带、磁盘、固态盘、非易失内存、主存、DNA等）
  - 低延迟/高吞吐量
  - 能效
  - 自治
  - 可扩展
  - 广域
  - 鲁棒性
  - 整合键值存储/数据库等
  - 数据全生命周期每比特极致性价比

COFS
China Open File System

谢谢！

高速、智能、开放：
中国开源文件系统（COFS）助力下一代存储系统的发展

曾令仿

新型智能计算系统研究中心

之江实验室

2020年10月