Technology and Application of Big Data

Qing LIAO(廖清)

School of Computer Science and Technology

HIT

Course Details

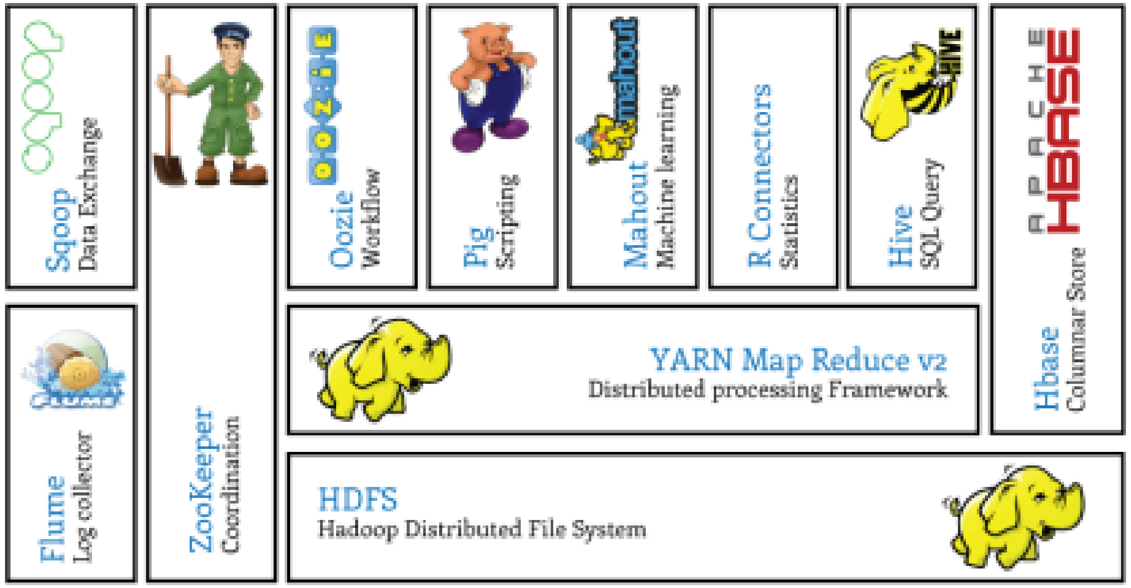
- Instructor:
    - Qing LIAO, liaoqing@hit.edu.cn
    - Rm. 303B, Building C
    - Office hours: by appointment
- Course web site:
    - liaoqing.me
- Reference books/materials:
    - Big data courses from University of California
    - Book: BIG DATA: A Revolution That Will Transform How We Live, Work, and Think
    - Papers
- Grading Scheme:
    - Paper Report 30%
    - Final Exam 70%
- Exam:
    - 21st July(Friday), 14:00-16:00, A502

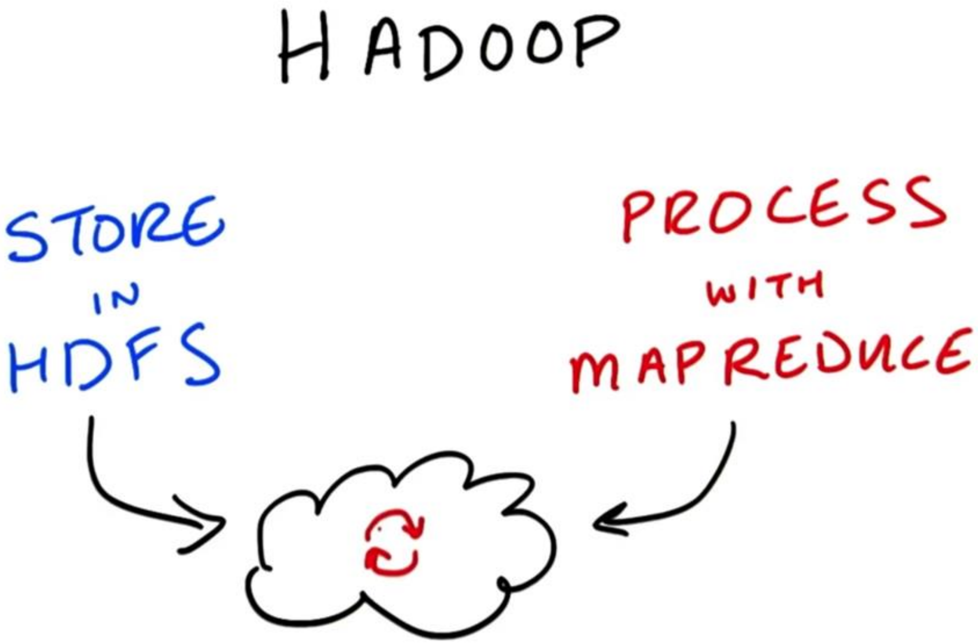
What You Learnt: Overview

- Topics:
  1) Introduction of Big Data
  2) Characterizes of Big Data
  3) How to Get Value from Big Data
  4) Technologies of Big Data
  5) Applications of Big Data
- Prerequisites
  - Statistics and Probability would help
    - But not necessary
  - Machine Learning would help
    - But not necessary

Previous Section

Hadoop Eco-System

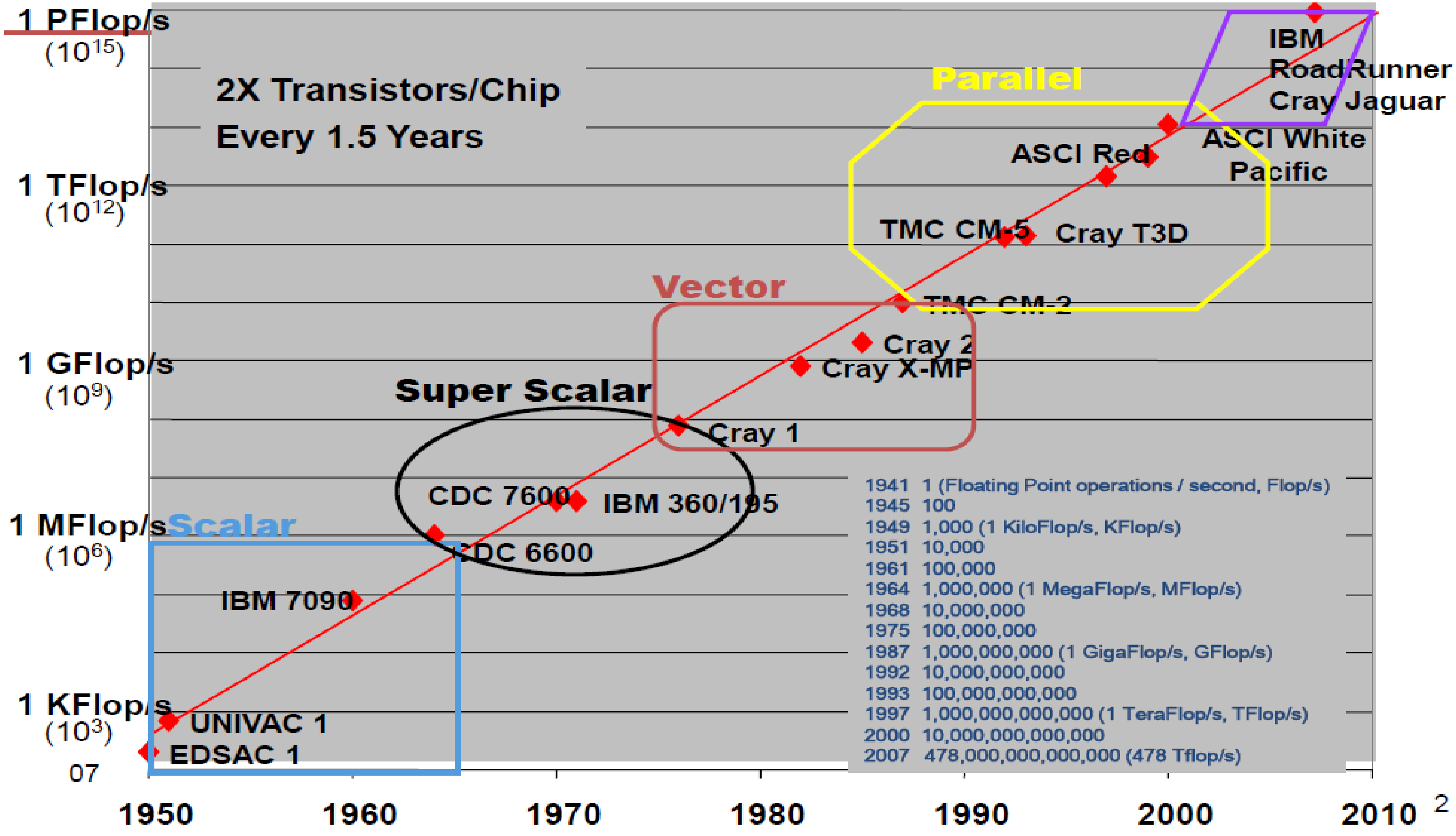




Supercomputer

- TianHe(Milk Way) 1 Supercomputer
- NO.1 in “The International Conference for High Performance Computing, Networking, Storage and Analysis(SC10)”
- 2010.11.18， USA

Supercomputer



1941	1 (Floating Point operations / second, Flop/s)
1945	100
1949	1,000 (1 KiloFlop/s, KFlop/s)
1951	10,000
1961	100,000
1964	1,000,000 (1 MegaFlop/s, MFlop/s)
1968	10,000,000
1975	100,000,000
1987	1,000,000,000 (1 GigaFlop/s, GFlop/s)
1992	10,000,000,000
1993	100,000,000,000
1997	1,000,000,000,000 (1 TeraFlop/s, TFlop/s)
2000	10,000,000,000,000
2007	478,000,000,000,000 (478 Tflop/s)

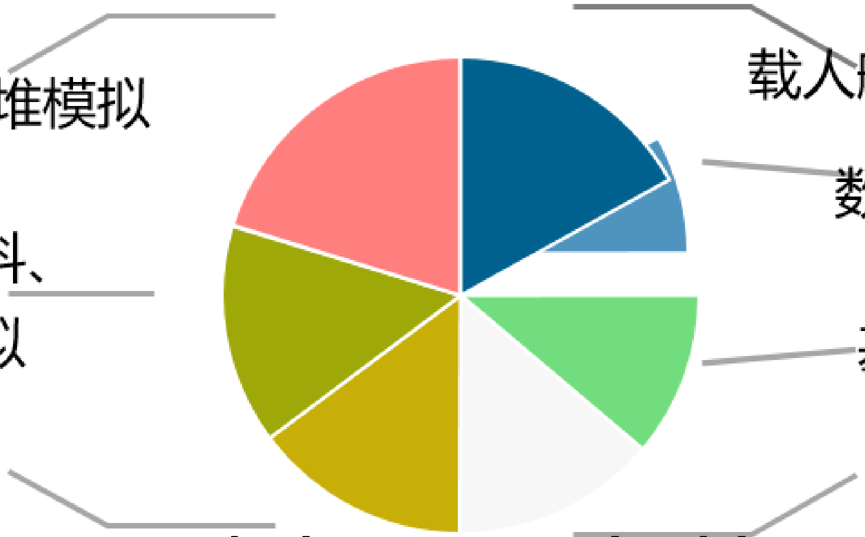
高性能计算机：当时计算处理能力最强大的计算机！

- 速度超级快、容量超级足、体积超级大、耗电超级多
- 超级计算机，巨型机

用途：预测和发现客观世界运动规律和演化特性的全过程

- 无损伤模拟真实实验无法进行的事情（海啸、核爆、气候等）
- 全过程全时空诊断，充分了解和细致认识研究对象
- 低成本短周期反复细致地进行

21世纪发展和保持核心竞争力的必需科技



核爆炸模拟、核反应堆模拟

能源勘探、新材料、
低温磁性材料模拟

大型装备设计制造

载人航天、深海探索、地球模拟

数值风洞、引擎燃烧3D模拟

基因工程、分子动力学模拟

密码工程 ……

Thanks NUDT Provides Slides

High Performance Computing-HPC

科学与工程计算：“挑战性”应用的“六超”特征

- 尺度超大（Too big）：宇宙模拟、地球模拟、互联网
- 尺度超小（Too small）：粒子物理、基因工程
- 时变超快（Too fast）：海啸、飓风、地震模拟
- 时变超慢（Too slow）：人类起源演变、气候变化预测
- 过程超危险（Too dangerous）：核爆炸模拟、核反应堆模拟
- 过程超昂贵（Too expensive）：大型风洞、汽车碰撞试验

Thanks NUDT Provides Slides

High Performance Computing-HPC

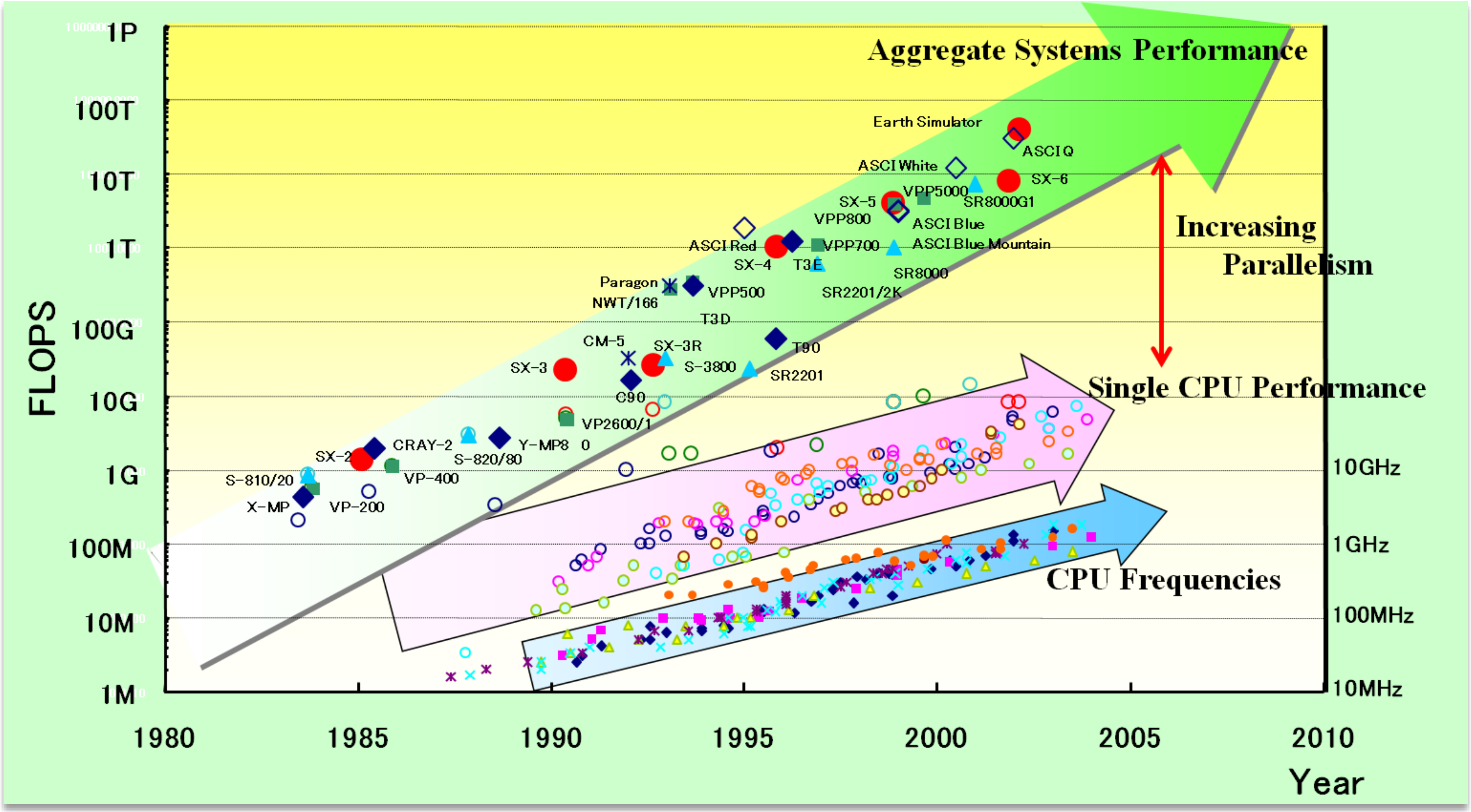
提高计算性能的“三驾马车”

- 提高主频，提高CPU性能
  - ➢ 因功耗及冷却制约，曾延续15年以上的按指数增长的主频，已渐趋停止。从2004年起就发生转折，一直保持在3－4GHz上下
- 优化结构，提高指令级并行及流水深度，提高CPU性能
  - ➢ 提升指令级并行（ILP）及深度流水技术潜力几已挖尽，导致结构复杂、功耗增加、得不偿失
  - ➢ AMD k8 采用12级深度流水线，每拍执行3条指令，运算部件仅占10%芯片面积
- 扩大并行度，提高全系统的性能

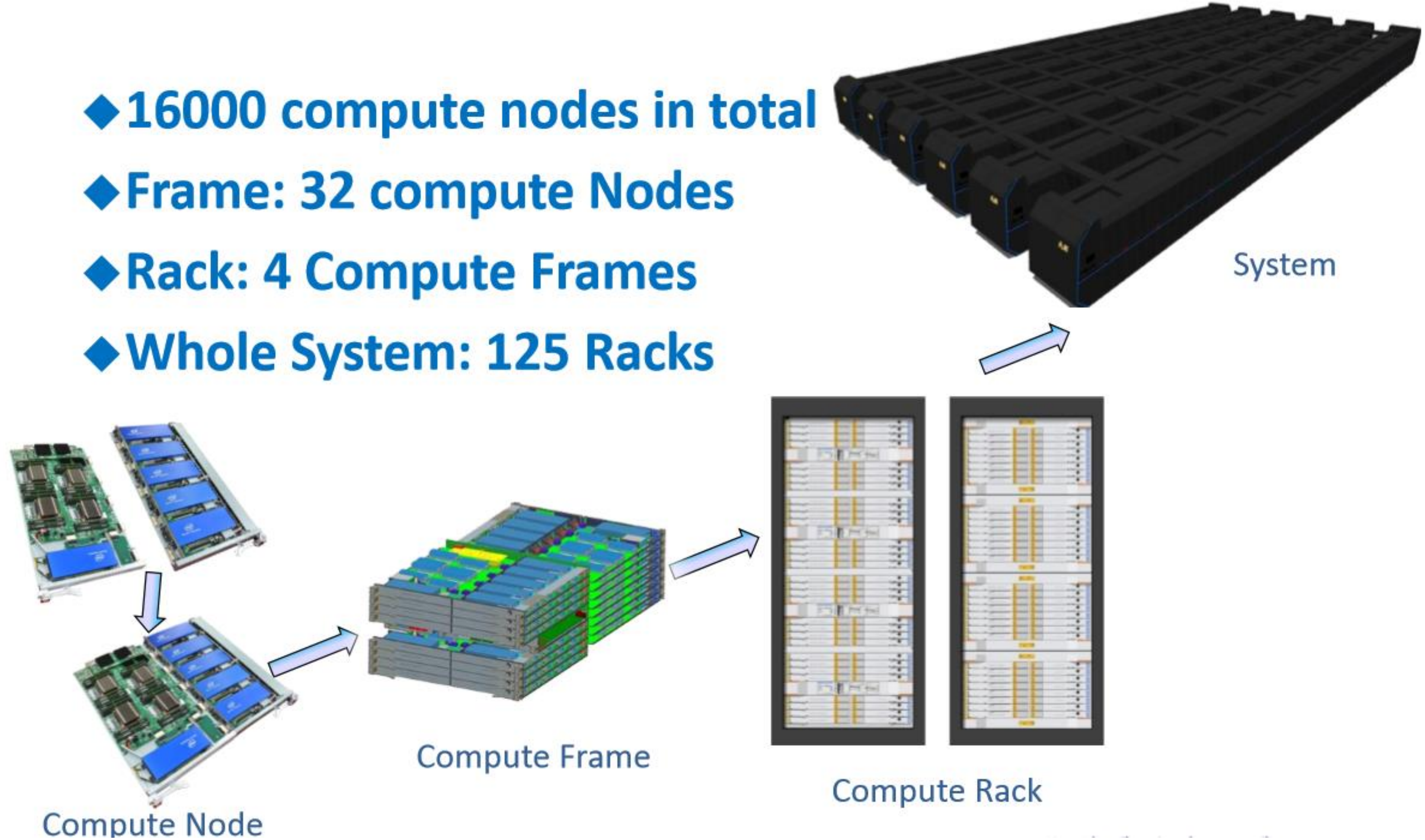
Thanks NUDT Provides Slides

High Performance Computing-HPC

“Performance = Parallelism”



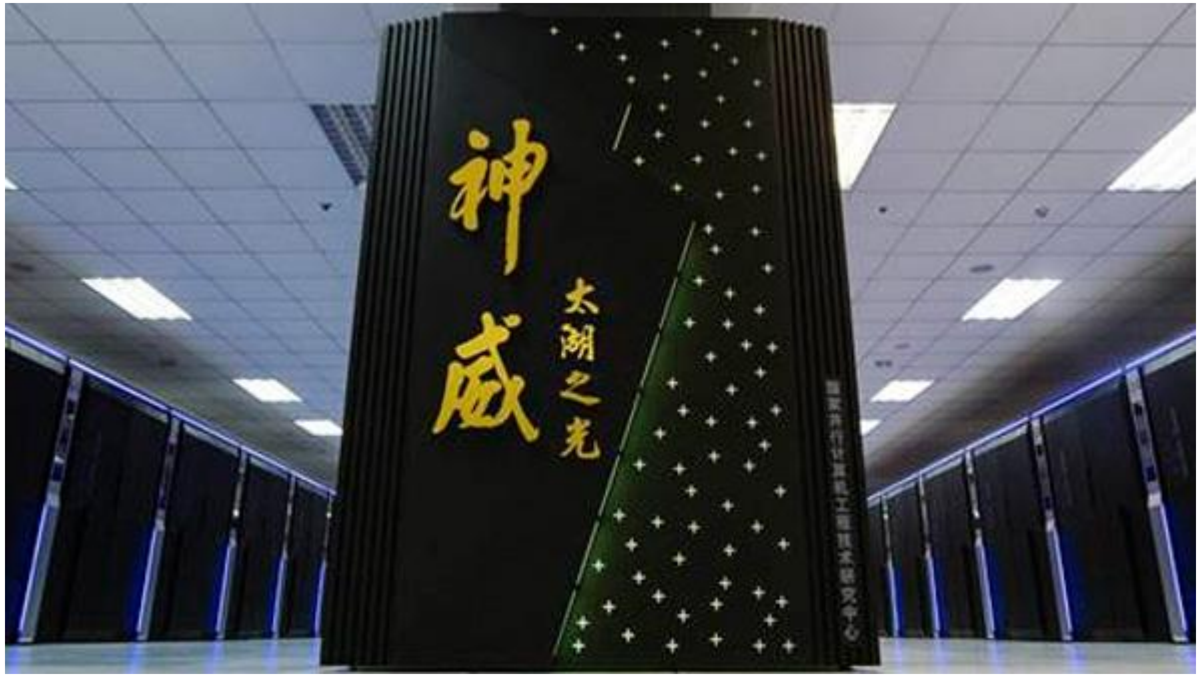
Supercomputer

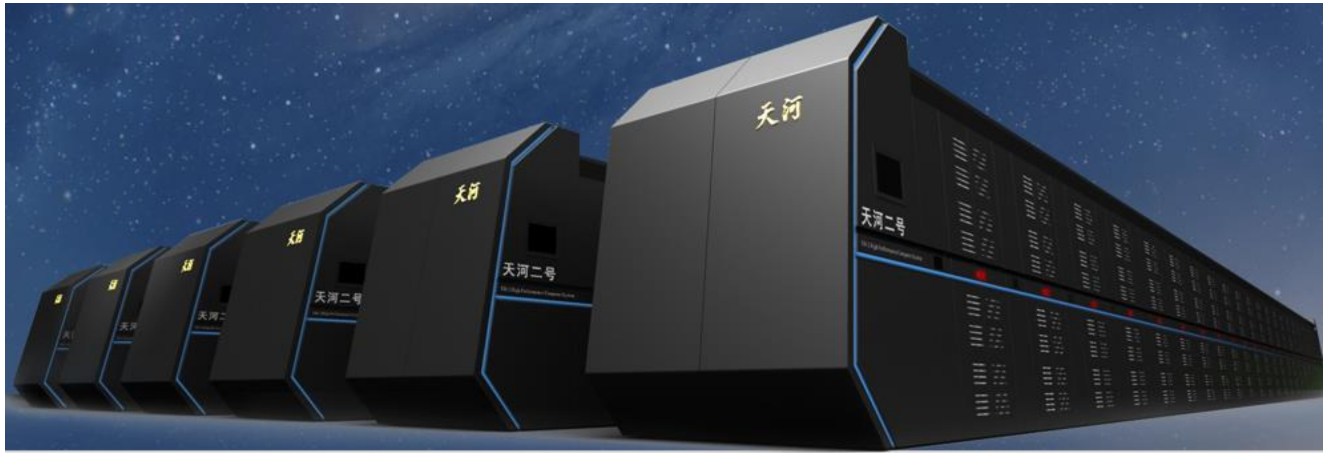


- 16000 compute nodes in total
- Frame: 32 compute Nodes
- Rack: 4 Compute Frames
- Whole System: 125 Racks

Top 500

2010,11	TianHe 1
2013,06	TianHe 2
2013,11	TianHe 2
2014,06	TianHe 2
2014,11	TianHe 2
2015,06	TianHe 2
2015,11	TianHe 2
2016,06	Sunway TaihuLight
2016,11	Sunway TaihuLight
2017,06	Sunway TaihuLight





Supercomputer

- NSCC-GZ Motivation
  - ➢~100 petaflops system
  - ➢863 High tech. Program of Chinese Government
  - ➢Government of Guangdong province and Government of Guangzhou city
- NSCC-GZ
  - ➢Open platform for research and education
  - ➢Public information infrastructure
- Goal
  - ➢Scalability
  - ➢Power consumption
  - ➢Resilience
  - ➢Usability