HPC Systems (in student cluster competitions)

GeekPie_HPC Tutorial #3

陈宸

HPC Systems

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	2,397,824	143,500.0	200,794.9	9,783
2	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM / NVIDIA / Mellanox	1,572,480	94,640.0	125,712.0	7,438
3	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
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61,444.5	100,678.7	18,482
5	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	387,872	21,230.0	27,154.3	2,384

HPC Systems At A Glance

Sunway TaihuLight

- 40960 SW26010 CPU
- 1.31PB RAM
- CPU for Computing
- 20PB storage
- Infiniband
- 15 MW on LINPACK
- 105 PFLOPS LINPACK
- RaiseOS 2.0.5

Our Mini HPC for SCC

- 10 Intel Gold 6132 CPU
- 1.28TB RAM
- 16 NVIDIA Tesla V100 GPU
- 800G shared storage
- Infiniband + Ethernet
- 3000W limit on LINPACK
- 41 TFLOPS LINPACK
- CentOS 7

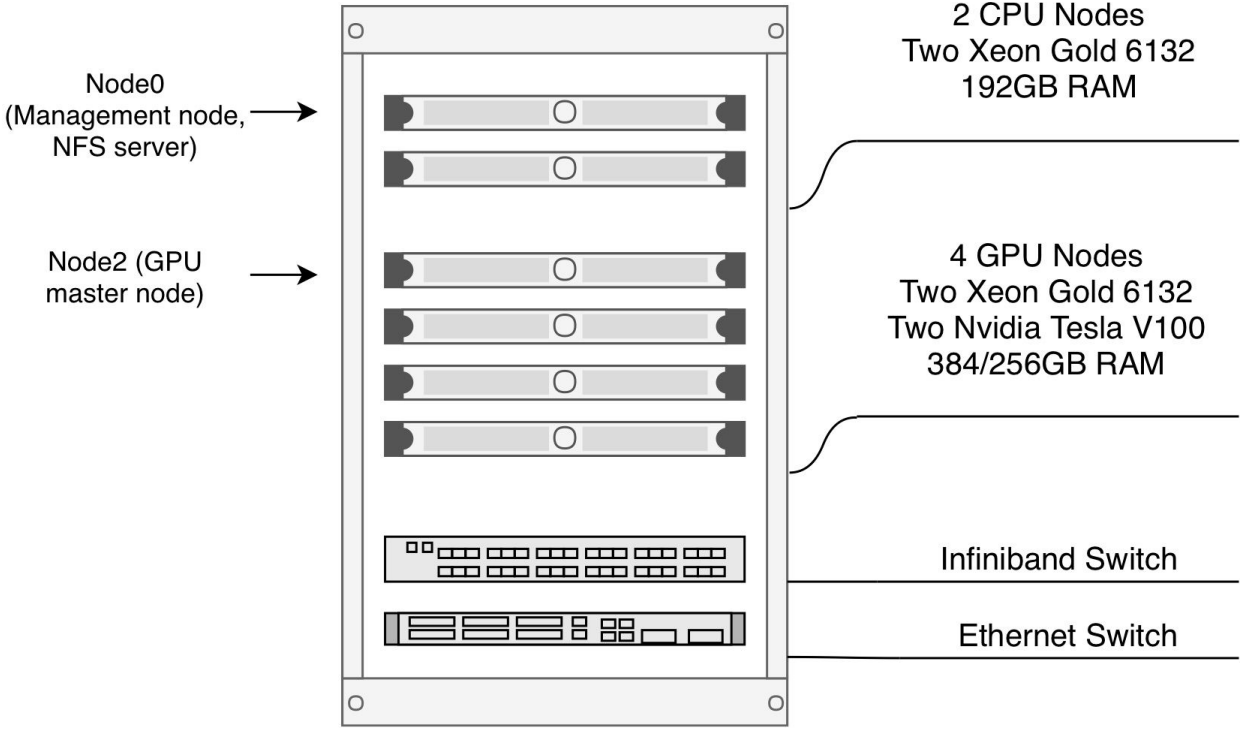
Large Scale

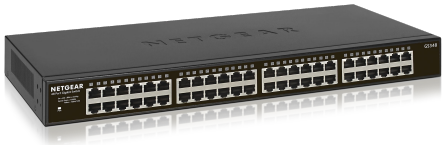
Small Scale

Components

- Single node setup
  - Computing unit (general and heterogeneous)
  - Primary storage (RAM)
  - Secondary storage (shared and local)
  - Power and cooling system
- Inter-node communication
  - Ethernet and Infiniband
  - Protocol (different layers)
  - Topology
- Per-node configuration
  - Operating system (choice, quick deployment)
  - Driver, file system, SELinux, firewall
  - Runtime library, compilation toolchain

Our Setup (ASC18)



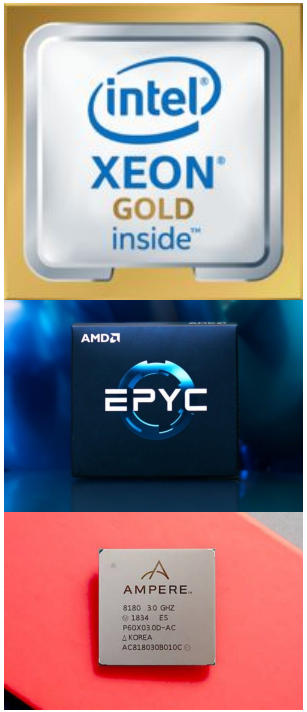


Making Decisions

- Fill all the DIMM slot, Dual Xeon Platinum, 8 Tesla V100 per node, all Intel DC SSDs, Infiniband EDR interconnections...?
- Yes, but...
  - Poor budget?
  - Power limit?
  - Chassis space?
- We have to make decisions, and gain trade-offs.
- It is hard, but luckily, we have only a few nodes to configure!
  - Typically, we only have to consider 2-8 nodes in student cluster competitions.

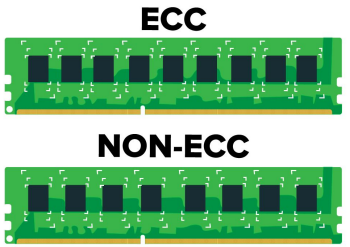
CPU - The Core

- Which ISA to choose?
  - X86 - General computing
  - ARM - RISC, power efficiency
  - Others...
- Compatibility matters
- X86_64 for now, maybe others for future
- PCIe lanes, RAM support should also be considered
- Application type
- Power consumption:
  - Intel Xeon Gold 6132, 14C28T, TDP 140W (CPU-B 27522)
  - Intel Xeon Platinum 8180, 28C56T, TDP 205W (CPU-B 37312)
  - AMD EPYC 7601, 32C64T, TDP 180W (CPU-B 33545)



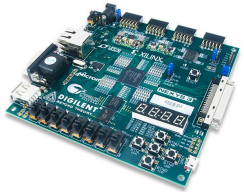
Storage

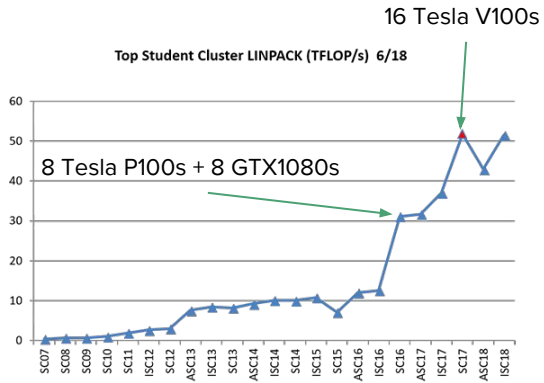
- Primary storage - RAM
  - Typically, a DDR4 16G stick costs < 10 watt
  - 64GB, 128GB, 192GB, 256GB... Trade-offs
  - Dual channel / quad channel
  - Always go with ECC
- Secondary storage - HDD, SSD
  - HDDs are out-of-date
    - Poor random performance
  - Enterprise-class SSDs are better
    - Intel DC SSDs, reliable
  - RAID for redundancy?
- NVM - A new trend
  - Byte-addressable, non-volatile, low latency, high speed



Heterogeneous Computing

- CPUs are great for general computing, but not optimized for specific workloads
- We need hardware acceleration!
- GPUs
  - NVIDIA GPUs rules SCC in recent years
  - The more, the better
  - Tricky power control
- Xeon Phi
  - Used to be popular years before
- ASIC or FPGA?
  - Possibly in the future...



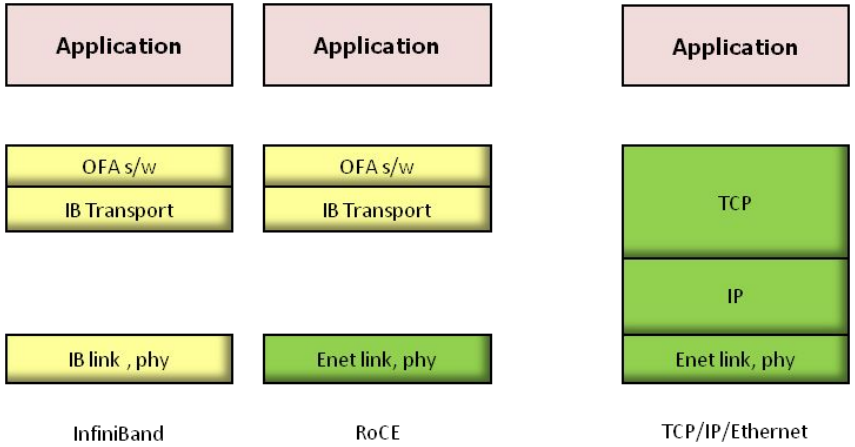


Power and Cooling System - Don’t Neglect

- Power supply
  - Stable and surplus
  - Redundant power supply
- Cooling system
  - A fan could consume up to tens of watts power
  - Protect other components
  - Monitor the system temperature distribution

Inter-node Communication

- Ethernet
  - Typically 1 Gbps or 10 Gbps
- Infiniband
  - Up to 100 Gbps
- Protocols
  - TCP/IP (the most commonly seen)
  - IP over IB
  - RDMA (Infiniband, RoCE)



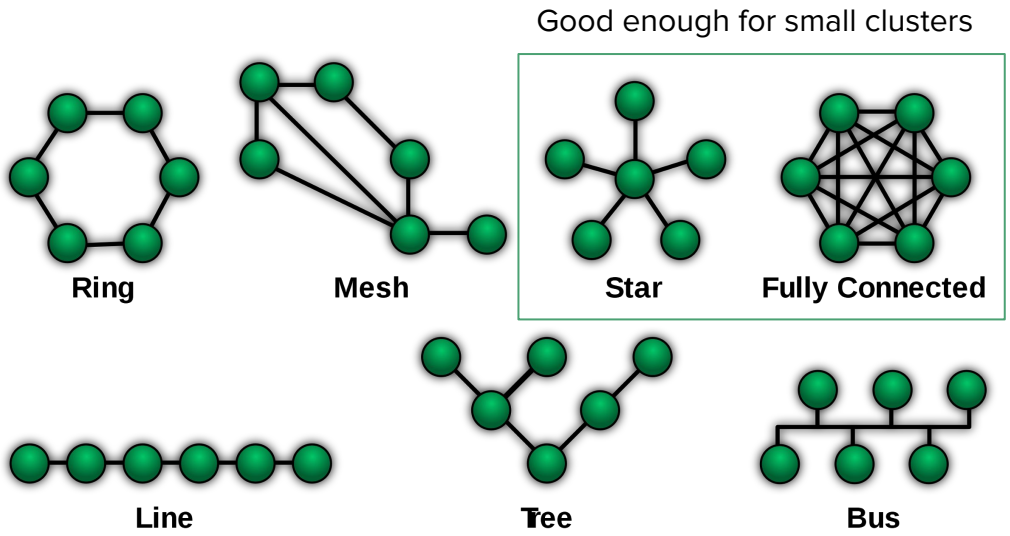
Links and Transceivers

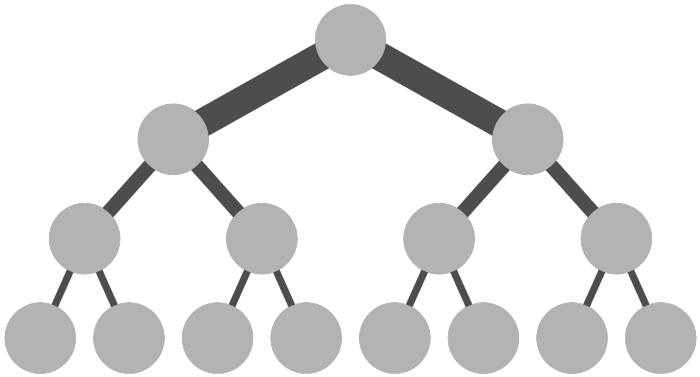
- IB Links

Characteristics									
	SDR	DDR	QDR	FDR10	FDR	EDR	HDR	NDR	XDR
Signaling rate (Gbit/s)	2.5	5	10	10.3125	14.0625[7]	25.78125	50	100	250
Theoretical effective throughput (Gb/s), per 1x[8]	2	4	8	10	13.64	25	50	100	250
Speeds for 4x links (Gbit/s)	8	16	32	40	54.54	100	200	400	1000
Speeds for 8x links (Gbit/s)	16	32	64	80	109.08	200	400	800	2000
Speeds for 12x links (Gbit/s)	24	48	96	120	163.64	300	600	1200	3000
Encoding (bits)	8/10	8/10	8/10	64/66	64/66	64/66	64/66	Undefined	Undefined
Adapter latency (microseconds)[9]	5	2.5	1.3	0.7	0.7	0.5	less?	Undefined	Undefined
Year[10]	2001, 2003	2005	2007	2011	2011	2014[8]	2017[8]	after 2020	future (after 2023?)

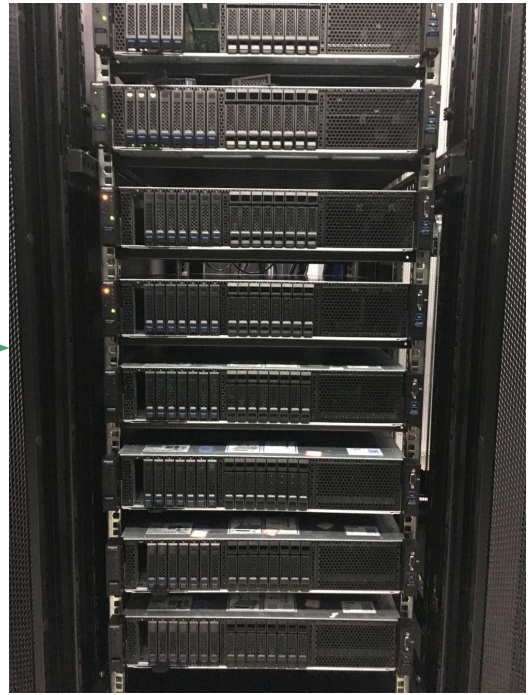
- Small form-factor pluggable transceiver (SFP)
  - 1 Gbit/s SFP (1000BASE, RJ45)
  - 10 Gbit/s SFP+ (10GBASE)
  - 40 Gbit/s QSFP+ (QDR, FDR10)
  - 50 Gbit/s QSFP14 (FDR)
  - 100 Gbit/s QSFP28 (EDR)
  - 200 Gbit/s QSFP56 (HDR)

Topology


Good enough for small clusters
Ring
Mesh
Star
Fully Connected
Line
Tree
Bus



Now we have set the machines up



Operating System Choice

- GNU/Linux, *BSD, Windows
- Generally we prefer GNU/Linux for compatibility, performance and familiarity
- Which distro?
- For SCC, I recommend CentOS
  - “Free” RHEL built for servers
  - Prioritized driver support
  - Sort of stable because it’s “stale”
- Why not Windows family?
  - Application developers don’t like Windows

Fast Deployment

- Even in SCC, we still have to deploy 4-8 machines.
- For consistency and efficiency, we need to deploy the operating system in a fast manner
- Option 1: disk clone
  - Configure once, clone to the rest
  - A useful tool: clonezilla (contained in archiso)
  - Not clean, but fast and fault-tolerant
- Option 2: PXE boot + automated installation tool
  - Setup a PXE server in LAN
  - Write automation scripts (like anaconda.cfg)
  - A clean installation is proceed

Shared and Local File Systems

- Since we have large memory, local file system performance is not a key factor
- Sophisticated file systems like xfs and ext4 are good enough
- Shared file system is necessary
  - For example, running a MPI program which requires all nodes to involve will let them load the same program at the same time
  - Therefore, a consistent global file system is needed
  - Set up NFS on a fast storage device, sharing libraries and programs
  - Configure NFS over RDMA
- Parallel file systems?
  - Necessary in large HPC data centers
  - Separated storage cluster
  - Not needed in SCC
    - Competition jobs are usually computing-intensive, not data-intensive

System Components

- Security
  - SELinux, firewall
- Drivers and libraries
  - GPU driver
  - IB driver (openib, manufacturer-provided)
  - CUDA toolkit
  - MPI (openmpi, mvapich, Intel MPI...)
  - Compiler (different versions and different languages)
  - Math (blas, lapack, Intel mkl...)
- User-level utilities
  - Power utilities (cpupower, nvidia-smi)
  - Shell utilities, tmux, screen...
  - Editors (vim, emacs...)
  - ...

Remote Access

- SSH
  - PubkeyAuthentication
- SSH X forwarding
  - Ssh -X
- VNC
- Tunneling under NAT
  - Frp (fast reverse proxy)
  - (Ideally) full-mesh P2P VPN like tinc (or zerotier)

Now we can test our system

- Basic:
  - Node communication (TCP, ICMP)
  - Proper MPI configuration
- Performance:
  - HPL
  - HPCG
  - HPCC
- Power:
  - Full-load power consumption
  - Power tuning
- Monitor
- Crash recovery

Done!

- Now we have our cluster set up
- What’s next?
  - Make && make install && ./run
  - Max out the cluster’s performance

- Next episodes:
  - HPC system profiling by 谢志强
  - HPC hardware platform in depth by 沈喆奇

Thanks!

GeekPie_HPC

扫一扫二维码，加入群聊。