Tier-1 Hardware Platform for the SPD experiment

Part 1 - computing

Kiryanov A.

VIII SPD Collaboration Meeting, 5-8 November 2024

Baselines

- No need for MPI, RDMA and other HPC stuff
- No need for GPUs and AI-optimized hardware
- We are mostly interested in performance and memory throughput **per core**
- AMD platforms are the most promising because of the memory subsystem performance
  - https://www.servethehome.com/memory-bandwidth-per-core-and-per-socket-for-intel-xeon-and-amd-epyc/
- Fat tree network topology, minimal inter-leaf traffic
- The goal is to provide 20 000 cores

An offer from HW supplier

- Compute server 2U
  - 2Twin2 (4 nodes), 4 x 128 cores (UP) @2.25GHz, 4 x 768GB RAM, 4 x (2x10GbE, 2x25GbE)
  - 12 DIMM slots, 12x64=768 GB, 128 cores per node (6 GB per core)
  - 2x2000W PSU, expected load >1500W per chassis
- Switches
  - Spine: 32x100GE QSFP28, 4x200GE QSFP56
  - Leaf: 48x25GE SFP28, 6x100GE QSFP28
- Rack 42U
  - 12 servers with 1U interval
  - TOR switch 1U
  - 6 144 cores per rack
- 4 racks total
  - 40 servers
  - Last rack is 1/3 full
  - Two spine switches (3U) in the last rack
  - Each spine switch needs 12x100G ports
  - 24 kW per rack
- Total power consumption ~80 kW
- Cost ~458 MRub

Suggested changes

- Compute server 2U
  - 2Twin2 (4 nodes), 4 x 96 cores (UP) @3.1GHz, 4 x 384GB RAM, 4 x (2x10GbE, 2x25GbE)
  - 12 DIMM slots, 12x32=384 GB, 96 cores per node (4 GB per core)
  - 2x2000W PSU, expected load >1500W per chassis
- Switches
  - Spine: 32x100GE QSFP28, 4x200GE QSFP56
  - Leaf: 48x25GE SFP28, 6x100GE QSFP28
  - We can shrink the number of ports but that only contributes ~1% of the total cost
- Rack 42U
  - 12 server with 1U interval
  - TOR switch 1U
  - 4 608 cores per rack
- 5 racks total
  - 53 servers
  - Last rack is 1/2 full
  - Two spine switches (3U) in the last rack
  - Each spine switch needs 15x100G ports
  - 24kW per rack
- Total power consumption ~106 kW
- Cost: 448 MRub (2% less expensive, 18% more performance per core)

Discussion time!

- Cost estimate is preliminary
  - Subject to 10% variations due to shipping costs, etc.
  - Assembly manpower is not included
  - Management network (IPMI) and switches are not included
- It is assumed that all engineering infrastructure (cooling, power, racks) is already there. At PNPI we have:
  - 12 free 42U racks
  - ~150kW of power and air cooling
- Part 2 – storage is yet to be estimated