Panda Networking Options
Ethernet vs. InfiniBand

Outline

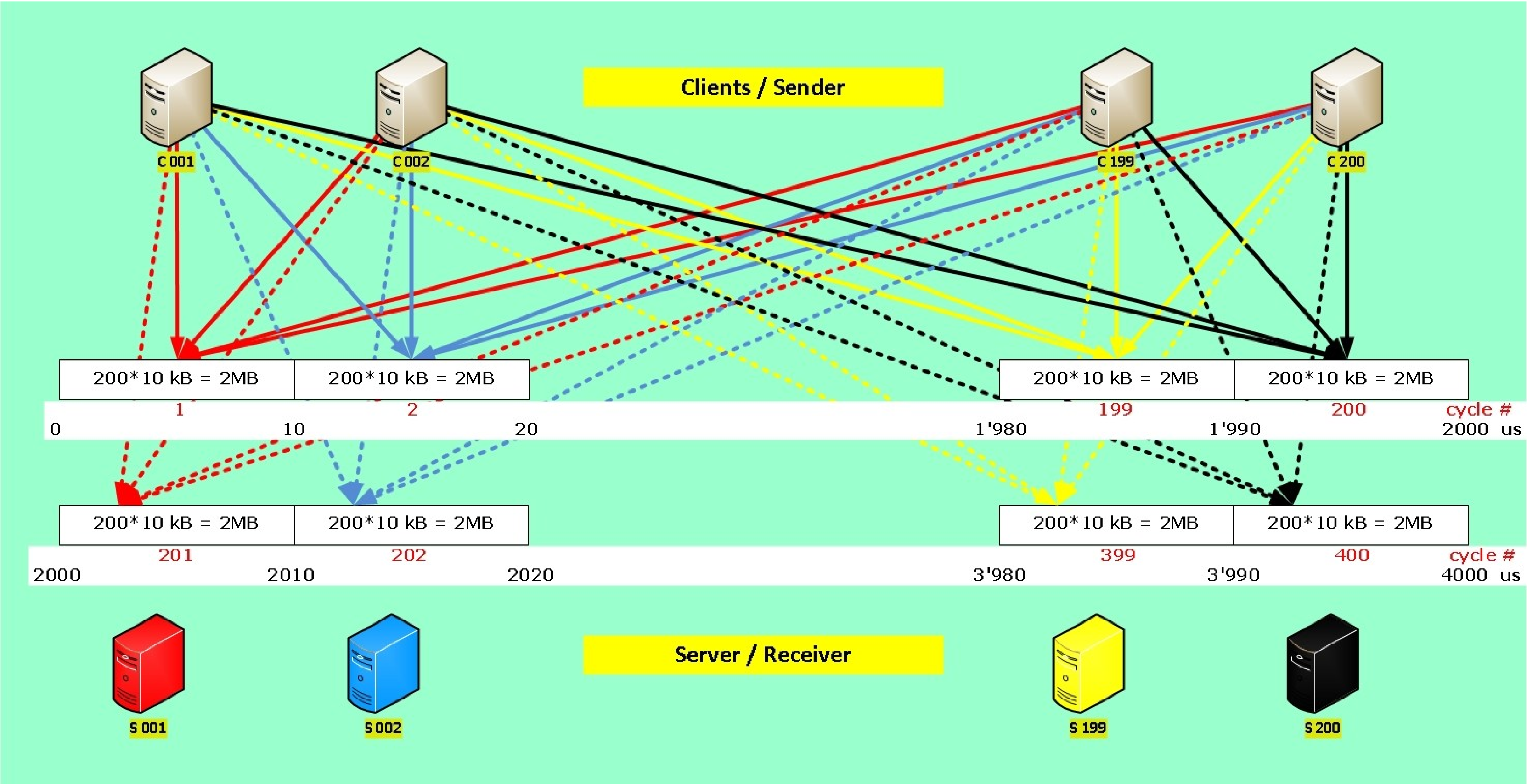
- Requirements
- Possible Ethernet Setup
- Possible InfiniBand Setup
- Ethernet vs. InfiniBand
- Conclusion

Network Requirements for PANDA

- 200 input streams with 10 Gb/s each
- 200 output streams with 10 Gb/s each
- Bursts of about 10 us
  - all sources send 10 kBytes to one destination
  - 10 us later all sources send to a different destination

- Consequences for worst case scenario:
  - 200 times 10 kBytes = 2 MBytes have to be stored in each input port

Requirements: II



- Gerhard Aker tried hard to explain this to equipment vendors

Ethernet

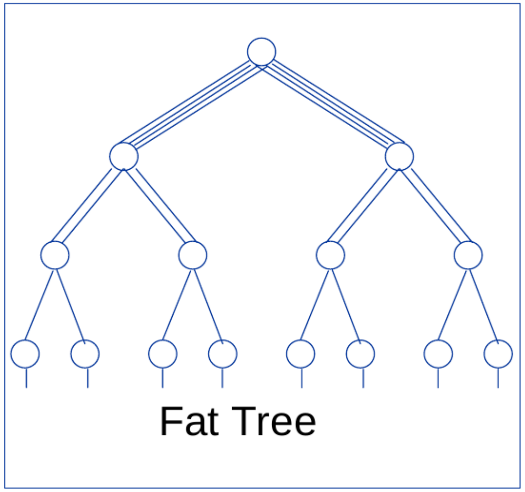
- Brocade and Arista have been contacted
  - Both claimed that this very special traffic is no problem for their switches
- Brocade answer:
  - Current model: 12 MBytes input buffers per port
  - Quote: 200k€ for 190x190 setup
- Aristas answer
  - 100 Mbytes/port buffer
  - Quote: 227k€ for 200x200 setup
- Cables:
  - 148€ / 10Gb/s cable (up to 7 m) => 50k€

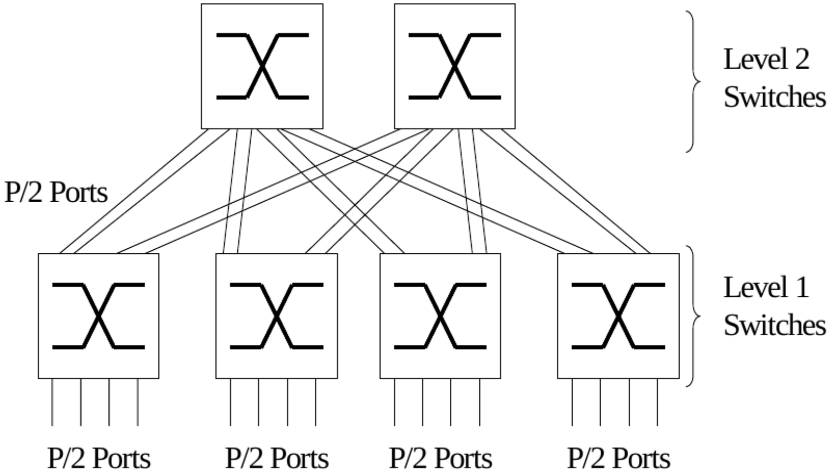
Ethernet II

- Looking into the future :-)
  - The GSI networking group deployed Ethernet networks from 1 Mb/s (pre-Ethernet) up to 10 Gb/s
  - Each new generation of Ethernet (normally a factor of 10 data rate increase) gave very roughly a factor of 3 reduction in price / (bit/s).
  - 10 Gb/s is mature now. 100 Gb/s existing but still expensive. Deployment of 100 Gb/s is expected soon.

Infiniband: I

- Possible PANDA-solution:
  - Fat-tree architecture
  - 9 InfiniBand switches (36 ports each) with 18 times 6 = 108 usable input/output ports, each nominal 40 Gb/s
  - Theoretical: 8Tb/s => more realistic: 2 Tb/s





- Pictures taken from: http://www.mellanox.com/pdf/whitepapers/IB_vs_Ethernet_Clustering_WP_100.pdf

Infiniband: II

- Price today for such a solution:
  - 45k€ switches + 30k€ cables => 75k€
- The disadvantage of this solution:
  - InfiniBand doesn't have significant input buffers (cut through switching used as low latency is the main concern for supercomputer clusters)
  - one needs traffic shaping/reordering (with memory)
    - Standard servers which also could do the transfer from Ethernet to InfiniBand
- Sergey Linev has shown that this works in large setups (774 nodes, 10Tb/s):
  - Up to 70% of the bandwidth is usable

Ethernet vs. InfiniBand: I

- **Optimized** InfiniBand is much cheaper:
  - ~75k€ vs. 250k€ for "plug&play" Ethernet
    - Comparable Ethernet solution: 130k€

- InfiniBand solution:
  - more complex and less reliable
    - Need many servers with traffic shaping software
    - Servers + software tend to be less reliable than plain switches
    - Additional costs for servers
  - "Single" vendor: Mellanox

Ethernet vs. InfiniBand: II

- Experience from the GSI-HPC group
  - “InfiniBand either works great or it doesn't. If it doesn't work it is hard to debug. Management is not simple!”
  - Bridging the Ethernet-world to InfiniBand for 1 Tb/s is in operation using normal servers with InfiniBand and Ethernet network adapters.
- Modern Ethernet switches have high level line speed diagnostics embedded
  - e.g. “Precision Data Analysis” from Arista
  - Today: external very expensive analysis hardware

Conclusion

- Networking solutions for PANDA:
  - Optimized InfiniBand is a factor of 2 cheaper than optimized Ethernet
  - “Plug&Play” Ethernet costs 250k€
  - InfiniBand setup is more complex and harder to maintain/debug
- Total costs for this mission critical (single point of failure) central part of PANDA is “small”
- It could even be considered to:
  - First use a solution for the needed bandwidths in the first years of operation and upgrade the network when needed
- Any commercial network has **many** huge advantages compared to a **highly** optimized custom solution: not even the price seems to be an argument!

Thank you for your attention!