Must. Go. Faster.

Selecting hardware for high-speed throughput measurement

Mark Feit · Internet2 / perfSONAR Development Team · mfeit@internet2.edu

perfSONAR is developed by a partnership of

Introduction

- *Must. Go. Faster.* …has been a recent theme.
- Why don't my servers go very fast?
- What do we need for 100 Gb/s? 200? 400?

How Fast Do You Need To Go?

- High speeds are great for link and device validation.
- Consumes network that users could be using.
  - May skew upgrade decision making
- Mathis *et al* applies the same at low speeds as high.
- Other measurements can help predict slowdowns.

What's Needed?

- A network card that can transmit and receive at "**full speed**"
  - The definition depends on your requirements
- A peripheral connection that can carry data at **full speed**
- Internal bus bandwidth to keep that connection fed at **full speed**
- Enough processors to make data to feed that bus at **full speed**

Peripheral Interconnect and Network Cards

Get on the Bus

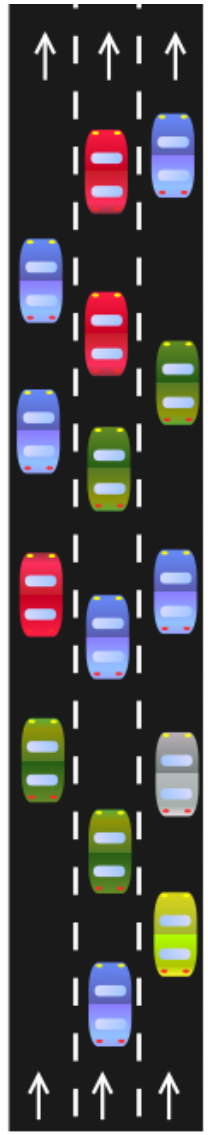
- **P**eripheral **C**omponent **I**nterconnect **E**xpress (PCIe)

- 2003 follow-on to 1992's PCI

Photo: Dmitry Nosachev CC-BY-SA 4.0

PCIe Lanes

- Classic PCI was 32- or 64-bit parallel.
- PCIe is multiple serial links called *lanes*.
  - Newer generations of PCIe use faster lanes.
- Switched architecture – Much less bus contention
- Slots and cards use "x" notation: x*N*, where *N* is the number of lanes.
  - Slot sizes: x1, x2, x4, x8, x16
  - Slower cards can be plugged into faster slots (e.g., x1 into x16)



The Large Print Giveth.

The Small Print Taketh Away.

- PCIe lanes are **full duplex**.
- Listed bandwidth is the aggregate of **both directions**.
- Single-direction bandwidth is **half** the PCIe advertised bandwidth.

One-Way Throughput Available Bandwidth

PCIe	Year	x1 Gb/s	x2 Gb/s	x4 Gb/s	x8 Gb/s	x16 Gb/s
1.0	2003	1	2	4	8	16
2.0	2007	2	4	8	16	32
3.0	2010	4	8	16	31	63
4.0	2017	8	16	31	63	126
5.0	2019	16	31	63	126	252
6.0	2022	30	60	121	242	484
7.0	2025	60	121	242	484	968
8.0	2028	121	242	484	968	1936

One-Way Throughput Available Bandwidth

PCIe	x16 Gb/s
3.0	63
4.0	126
5.0	252
6.0	484
7.0	968

PCIe 6.0 won't appear in servers until mid-2026.

PCIe 6/7 Compliance Tester
Summer, 2025

Cheat Codes for Slow Slots

- Some card manufacturers can connect one card to two slots
- NVIDIA (Mellanox) has Socket Direct
  - Doubles the available lanes on ConnectX-6 and later

Photo: NVIDIA

Cheat Codes for Slow Slots

PCIe	x32 Gb/s
3.0	126
4.0	252
5.0	484
6.0	968

⇦ Practical limit with today's servers

There's Always a Catch

- Everything must be arranged properly:
  - Cores
  - Bridge
  - Slots
  - Software

The Computer

Internal Bus Bandwidth

- Most systems that have PCIe slots can saturate at least one of the fastest of them.
- Old Raspberry Pis had a bandwidth problem
  - 1 Gb/s NIC but not enough internal bandwidth to go that quickly
  - Limit gone in RPi 4+
- Trust but verify
  - Test a couple of candidate systems before buying in quantity

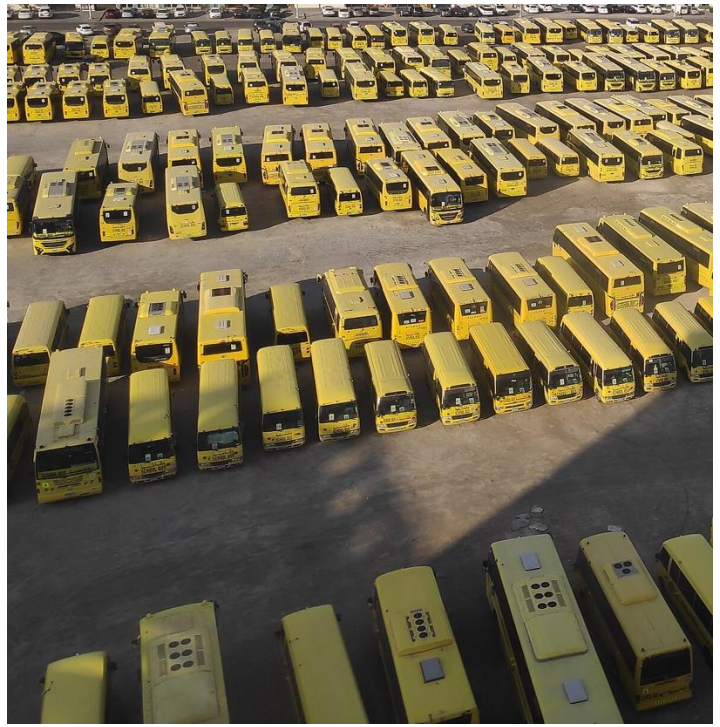


Photo: Gurav Dhwaj Khadka CC-BY-SA 4.0

Processors

- Don’t skimp on the CPU.
- 3+ GHz is known to work well.
- Cores? See the chart.
- Don’t forget to account for other workloads.

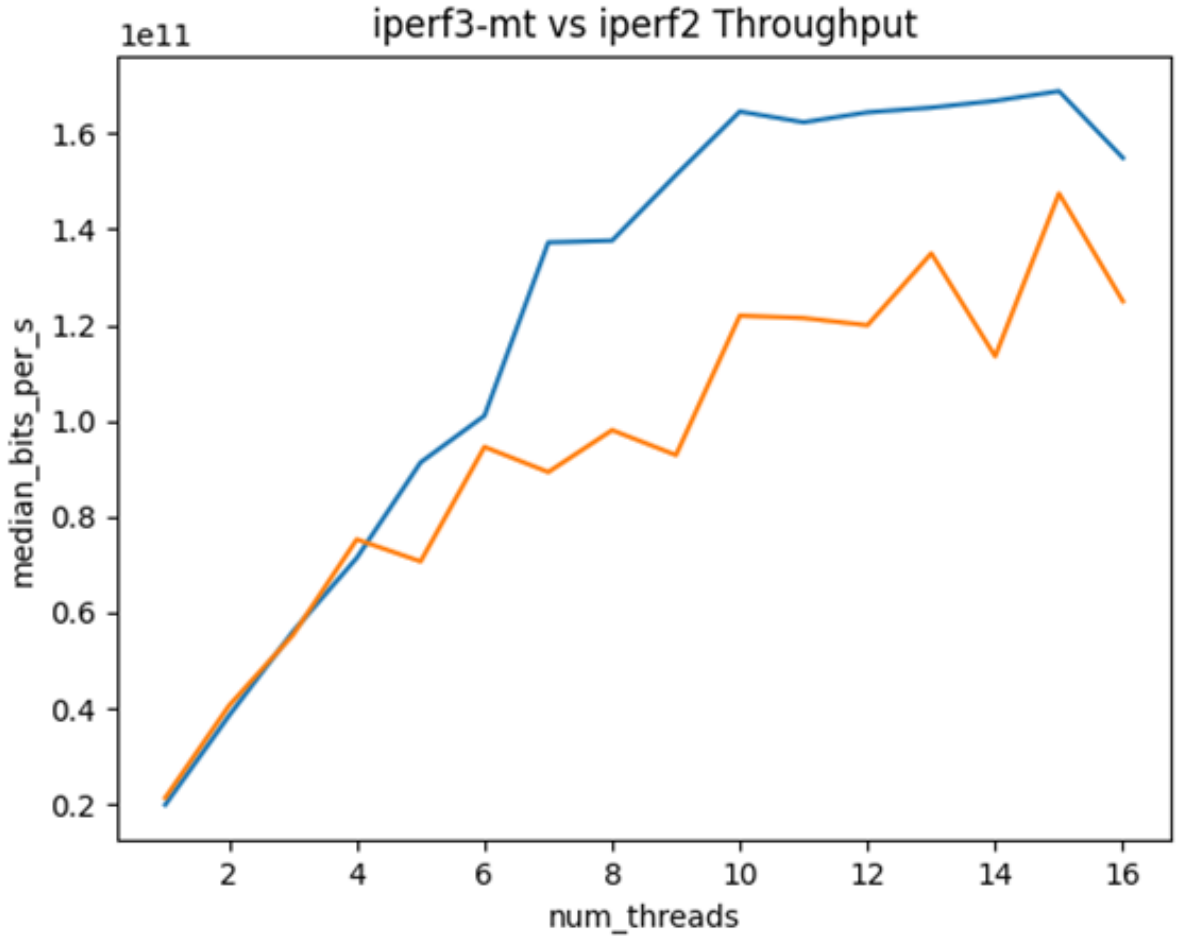


Chart: Sarah Larsen @ ESnet

Memory

- Large quantities aren't critical.
- Same eggs, more baskets:
  - More DIMMs improve DMA performance (not linear).
  - 1x 32 GiB < 2x 16 GiB < 4x 8 GiB
- Disable swappiness in the kernel.

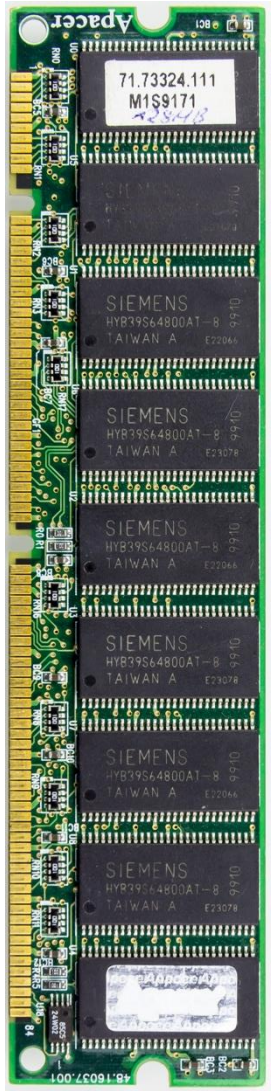


Photo: Raimond Spekking CC-BY-SA 4.0

Disk

- Not really a factor.

Photo: Evan-Amos CC-BY-SA 4.0

A Word About Link Validation

- Got glass? Rent dedicated hardware.
  - Better picture of the link
  - Measurements that software on a PC doesn't do
  - Waters muddied by L2/L3/system

Some Final Thoughts

- Throughput measurements are cool but not all-important on a live network.
  - Lower speeds can reveal many of the same problems
- Buy machinery that can support current needs with room to grow.
- Verify that proposed configurations can do the measurements you want **in the lab** before buying in quantity.

Questions?

Email

mfeit@internet2.edu

Question and answer icon by iconosphere from The Noun Project