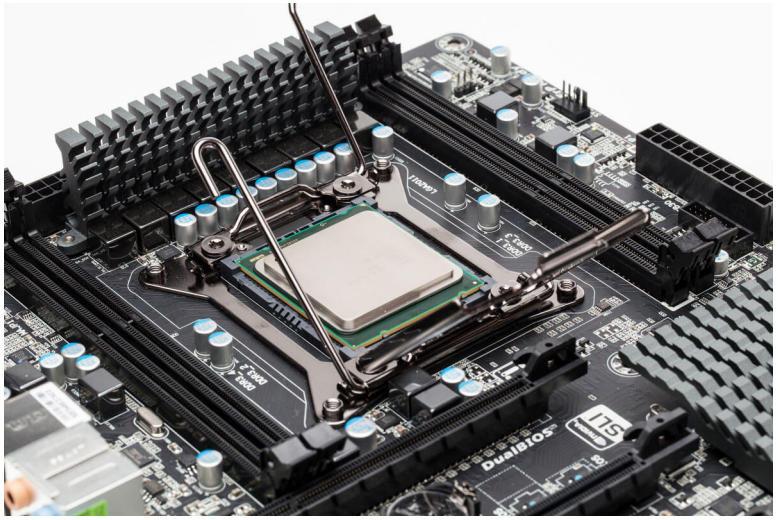


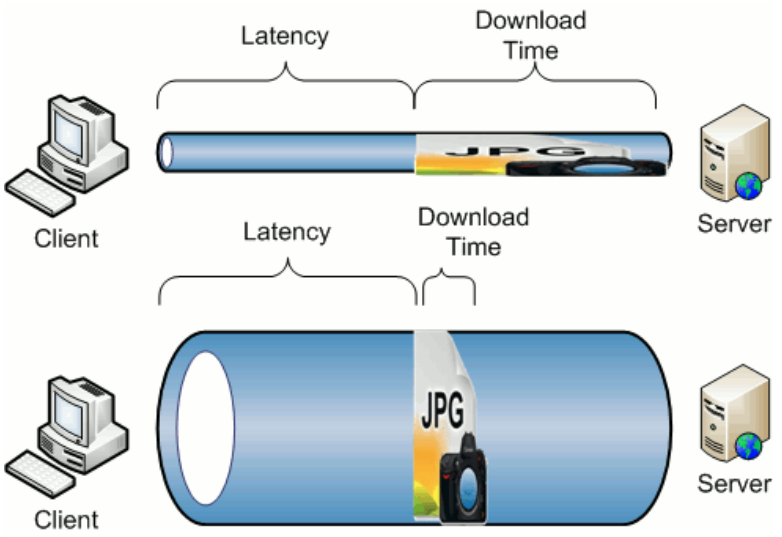
10: Computer Latency at Human Scale

https://github.com/matthiaskoenig/itbtechtalks

Dr Matthias König
Humboldt University Berlin,
Institute for Theoretical Biology

Important terms

- **bandwidth** is **maximum rate of transfer**, i.e., how much information can be send per time
- **latency** is **delay** until information reaches the destination
- **bits & bytes**
  - **bit** is the **smallest unit of storage** (stores 0 or 1)
  - **byte** is **collection of 8 bits** (2^8 = 256 states)
    - Kilobyte (KB), about 1 thousand bytes
    - Megabyte (MB), about 1 million bytes
    - Gigabyte, GB, about 1 billion bytes
  - **Storage in bytes, information transfer in bits**



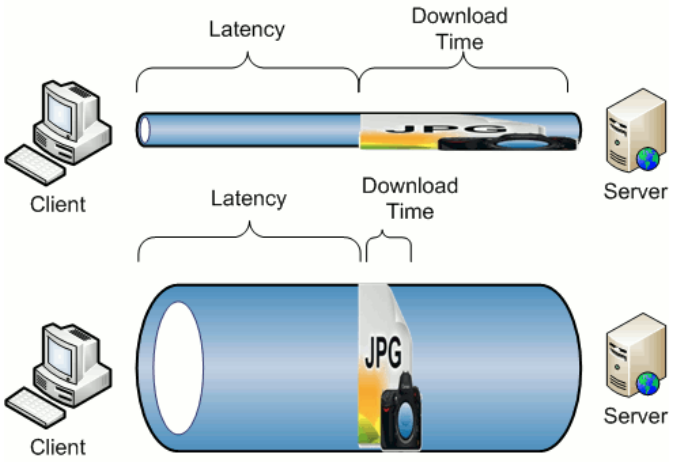
Ihr Ergebnis

Download-Geschwindigkeit:
907.018 kbit/s

Upload-Geschwindigkeit:
289.028 kbit/s

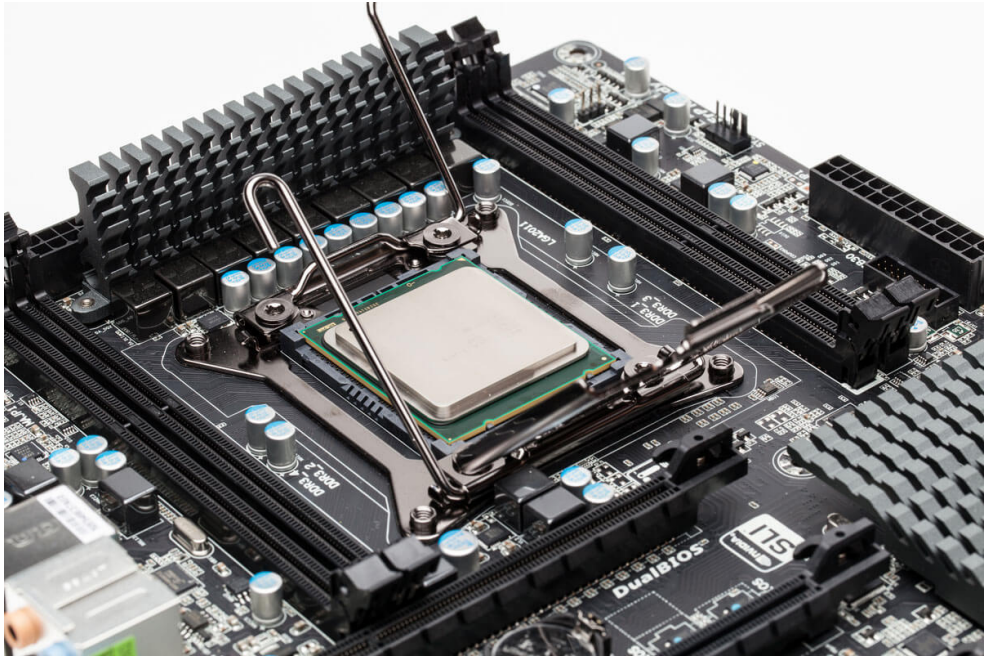
Ping-Geschwindigkeit:
24 ms

Perfekt Gut Befriedigend Zu gering



https://www.wieistmeineip.de

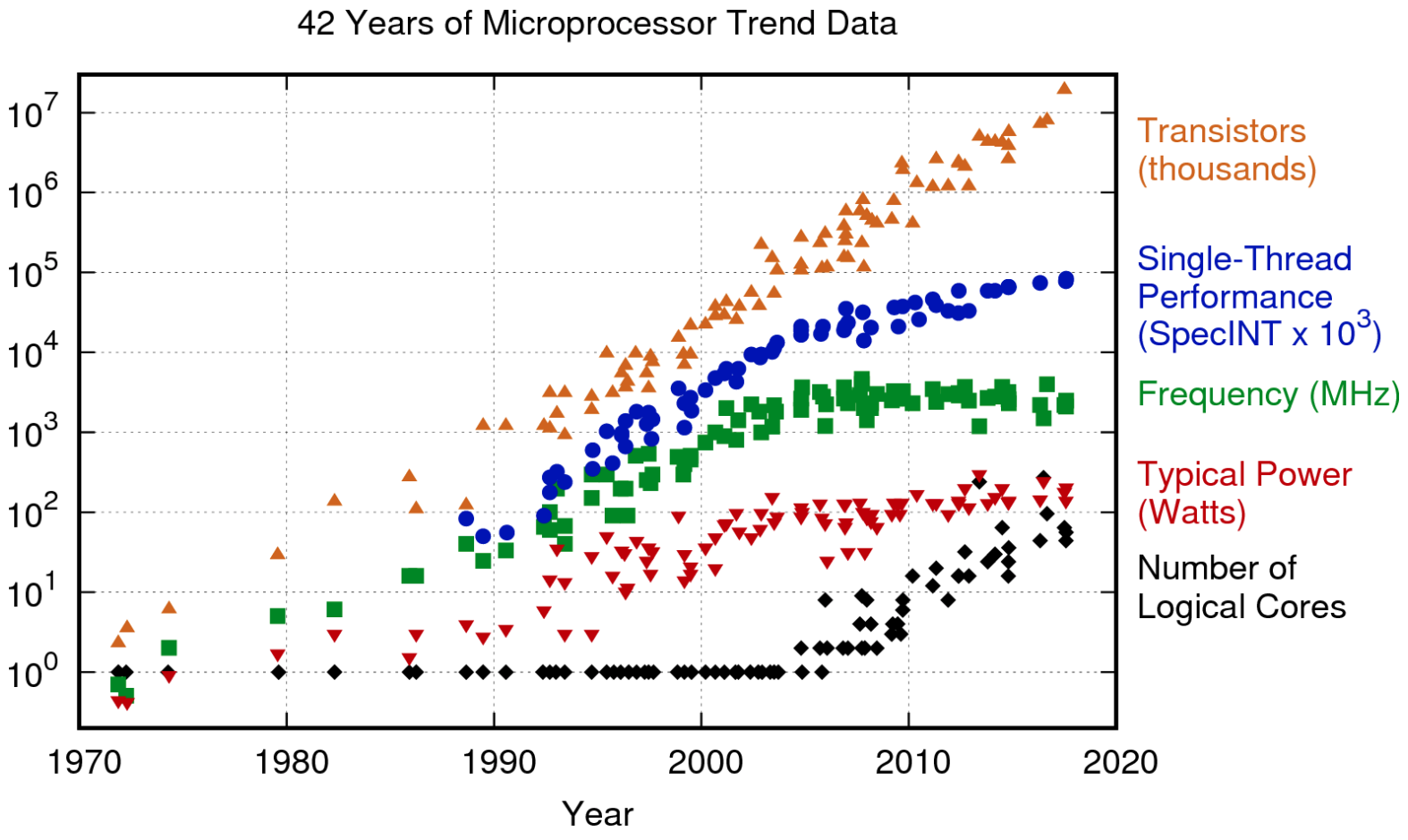
How fast are computer systems really?



3 000 000 000

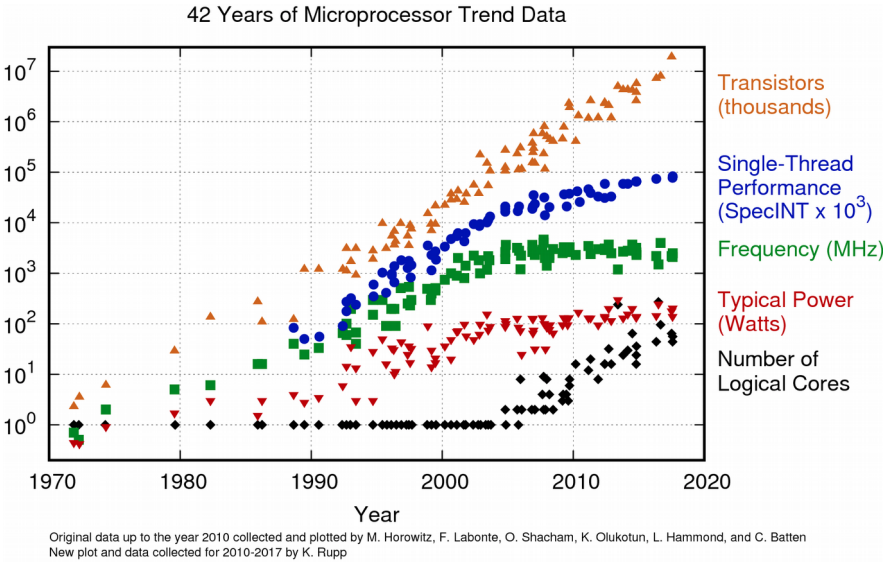
operations/second (3 GHz)

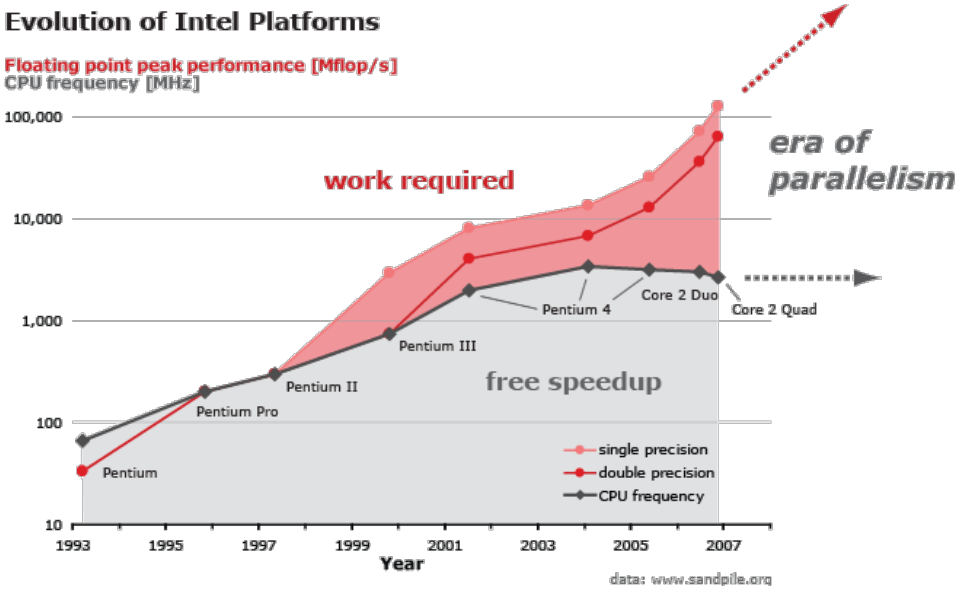
But no progress in the last 15 years?



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

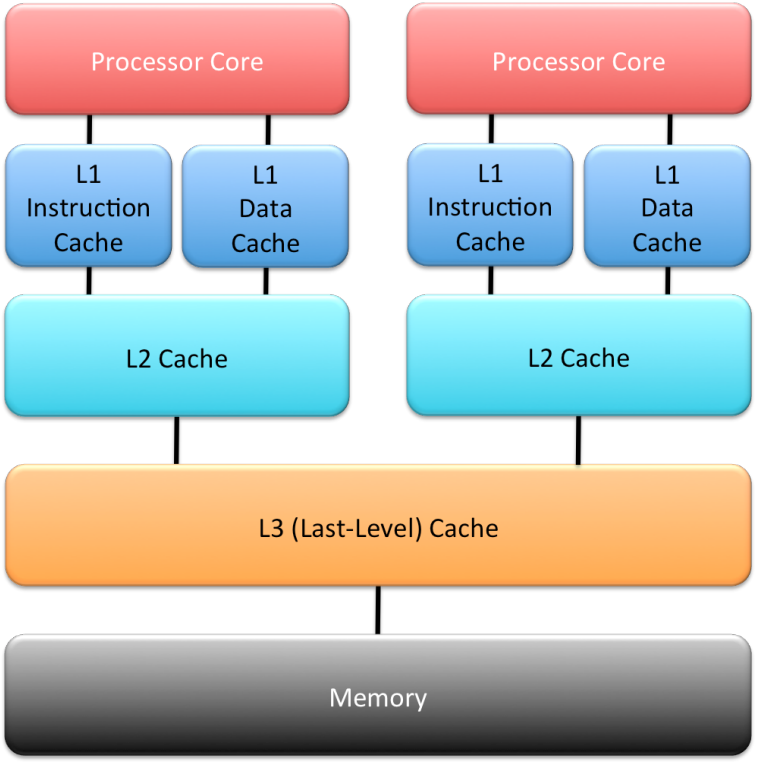
The future is parallel

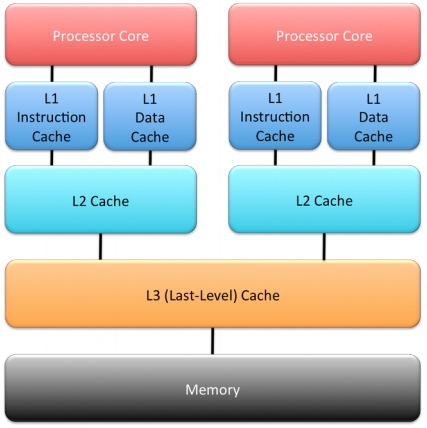




How does the data get into your CPU?

- A **CPU cycle** refers to a single tick of a processor's internal clock **~0.5ns**
  It is during the ticks of this clock that processors work their way through the pipeline of instructions awaiting computation.
- **Level 1 (L1) cache** directly connects to a CPU core, taking **~1ns**
- **Level 2 (L2)** cache takes about **3-6ns**
- Accessing **DRAM** takes anywhere from **60-100ns**

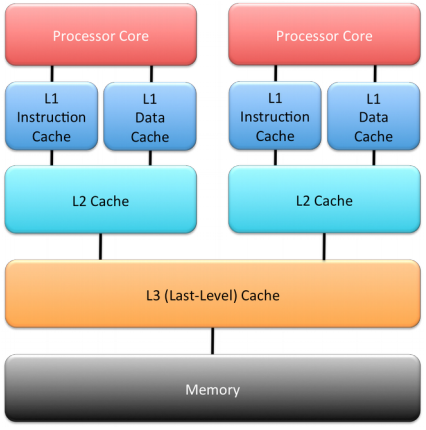




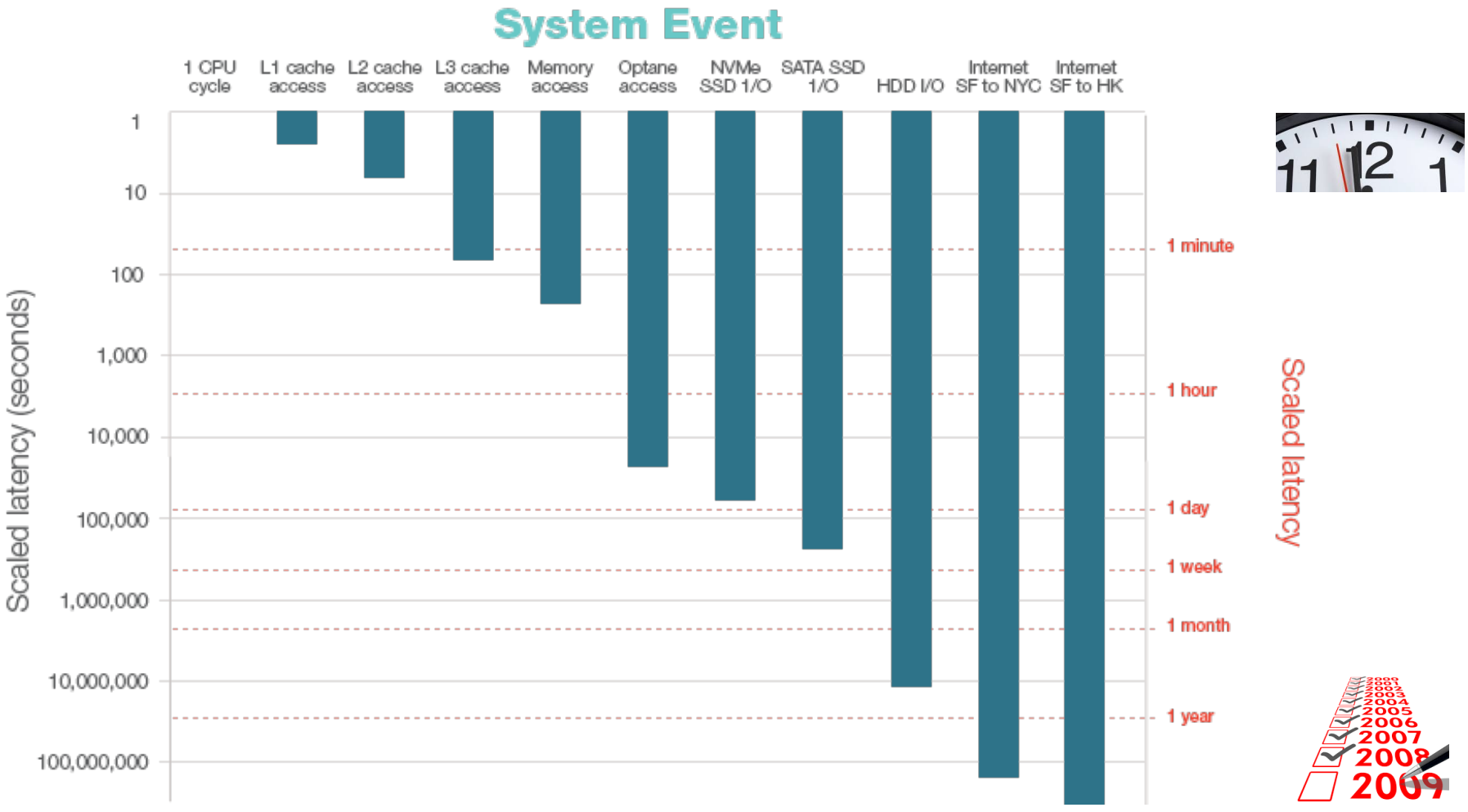
System Event	Actual Latency
One CPU cycle	0.4 ns
Level 1 cache access	0.9 ns
Level 2 cache access	2.8 ns
Level 3 cache access	28 ns
Main memory access (DDR DIMM)	~100 ns
Intel Optane memory access	<10 µs
NVMe SSD I/O	~25 µs
SSD I/O	50–150 µs
Rotational disk I/O	1–10 ms
Internet call: San Francisco to New York City	65 ms[3]
Internet call: San Francisco to Hong Kong	141 ms[3]

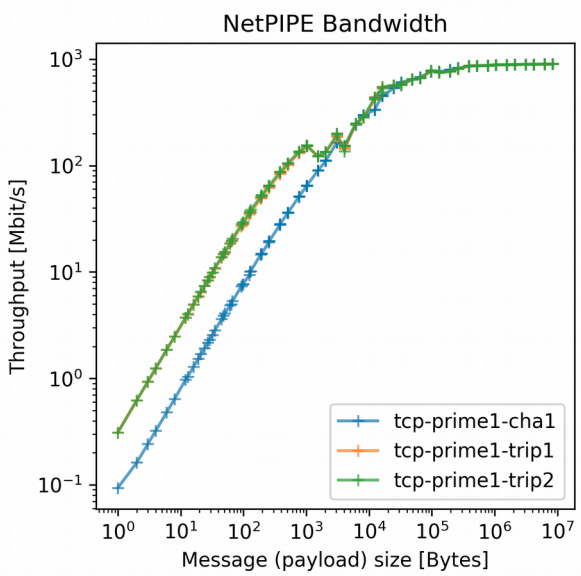
Latencies

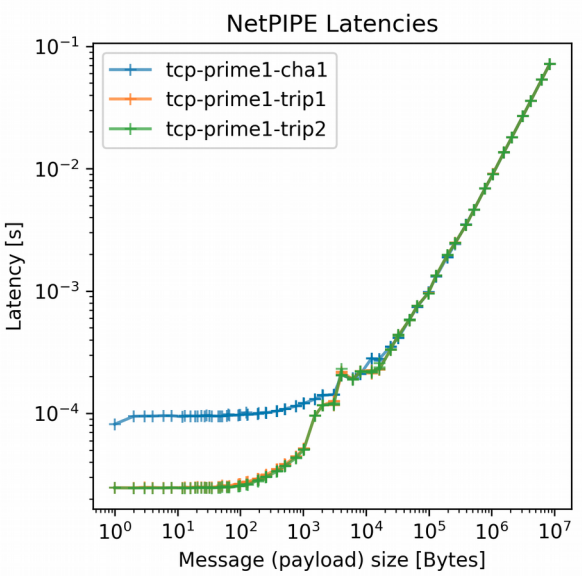
- As we move farther for the CPU, latency rises, but it doesn't do so smoothly.
- Moving from memory to storage is a huge performance hit.
- The move from SSD storage to spinning-disk is likewise huge—as is moving from disks of any kind to Internet calls.
- Rule of thumb: "**Always put storage/memory as close to the CPU as possible**"
- Numbers correct within order of magnitude

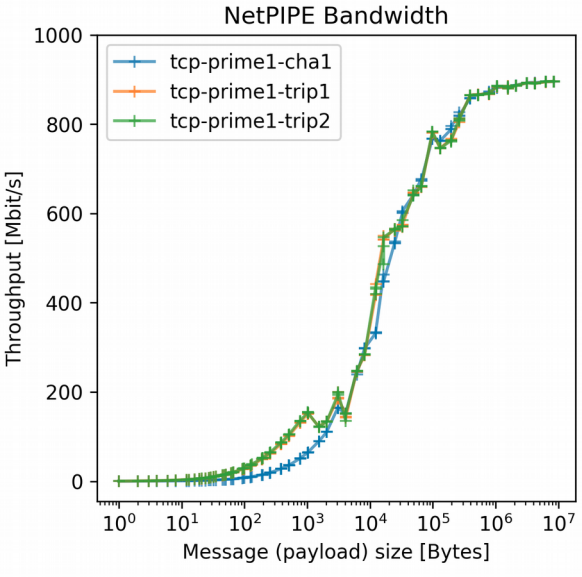


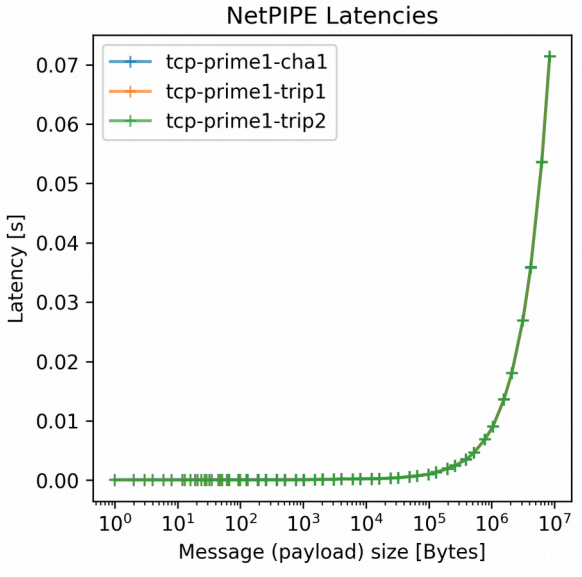
System Event	Actual Latency	Scaled Latency
One CPU cycle	0.4 ns	**1 s**
Level 1 cache access	0.9 ns	**2 s**
Level 2 cache access	2.8 ns	**7 s**
Level 3 cache access	28 ns	**1 min**
Main memory access (DDR DIMM)	~100 ns	**4 min**
Intel Optane memory access	<10 µs	**7 hrs**
NVMe SSD I/O	~25 µs	**17 hrs**
SSD I/O	50–150 µs	**1.5–4 days**
Rotational disk I/O	1–10 ms	**1–9 months**
Internet call: San Francisco to New York City	65 ms[3]	**5 years**
Internet call: San Francisco to Hong Kong	141 ms[3]	**11 years**


System Event
1 CPU cycle
L1 cache access
L2 cache access
L3 cache access
Memory access
Optane access
NVMe SSD 1/O
SATA SSD 1/O
HDD I/O
Internet SF to NYC
Internet SF to HK
Scaled latency (seconds)
1
10
100
1,000
10,000
100,000
1,000,000
10,000,000
100,000,000
1 minute
1 hour
1 day
1 week
1 month
1 year
Scaled latency









Latency vs. message size

- Ping-pong bandwidth and latency comparison (max bandwitdth 1Gbit/s ~ 125Mbyte/s)
  - cha1 (old hardware)
  - trip1/2 (new hardware)

- latency increases with bandwidth usage “**wait in line effect**”
- other factors are “**priority & scheduling**”, “**payload**”, “**protocol overhead**”, “**translation & serialization costs**”, “**speed of light**”
- “**Improvements made anywhere besides the bottleneck are an illusion**.”

References

- Computer latency at a Human scale https://www.prowesscorp.com/computer-latency-at-a-human-scale/
- Gregg, Brendan. “Systems Performance: Enterprise and the Cloud.” March 2015. www.brendangregg.com/sysperfbook.html. A CPU cycle refers to a single tick of a processor’s internal clock. It is during the ticks of this clock that processors work their way through the pipeline of instructions awaiting computation.
- Life of a Storage Packet (Walk), https://www.brighttalk.com/webcast/663/169543
- Some modifications in this table are based on: Intel. “Memory Performance in a Nutshell.” June 2016. https://software.intel.com/en-us/articles/memory-performance-in-a-nutshell.
- AT&T. “Network Latency.” May 2017. http://ipnetwork.bgtmo.ip.att.net/pws/network_delay.html.
- Hat tip to Nick Humrich’s blog post “Yes, Python is Slow, and I Don’t Care” for this particular example.