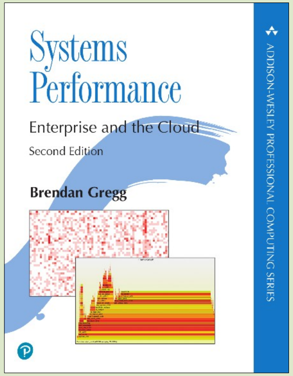
# Computing Performance 2022

## What's On the Horizon

Brendan Gregg

SRE CON ASIA PACIFIC

**SYDNEY, AUSTRALIA**
7–9 December, 2022



# Statement from the heart

I'd like to begin by acknowledging the Traditional Owners of this land and pay my respects to Elders past and present.

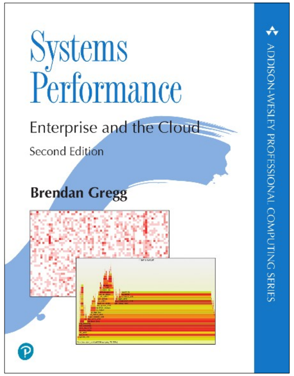
# Disclaimers: About this talk

This is

- **a performance engineer and author's views about server performance**

This isn't

- necessarily about my employer, my employer’s views, or USENIX's views
- an endorsement of any company/product or sponsored by anyone
- professional market predictions (various companies sell such reports)
- based on confidential materials
- necessarily correct or fit for any purpose




**My predictions may be wrong! They will be thought-provoking.**

# Agenda

1. Processors
2. Memory
3. Disks
4. Networking
5. Kernels
6. Hypervisors
7. Observability
8. AI

Not covering: Languages/runtimes, databases, file systems, front-end, mobile, desktop.

# Take Aways

- Awareness of current and future perf technologies
- Design faster systems to meet SLOs and performance needs
- Begin planning new technology support and maintenance

Slides: https://www.brendangregg.com/Slides/SREcon2022_ComputingPerformance

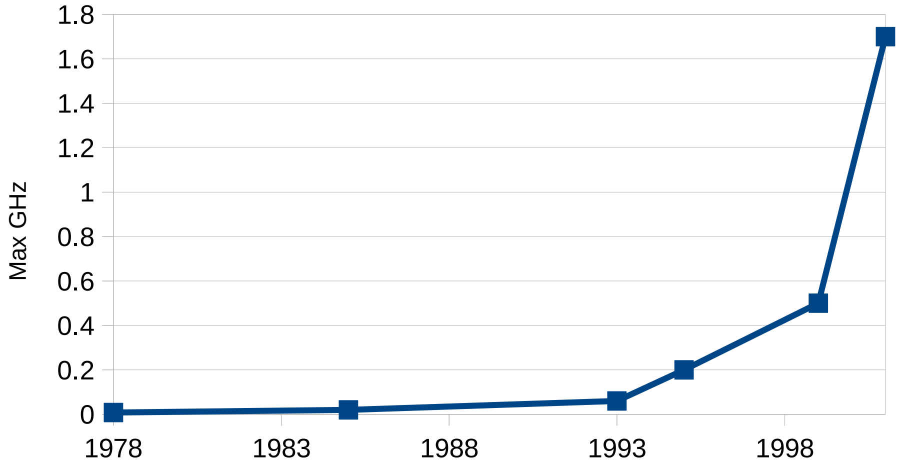
These contain extra footnotes as fine print!

# 1. Processors

# Clock rate

Early Intel Processors

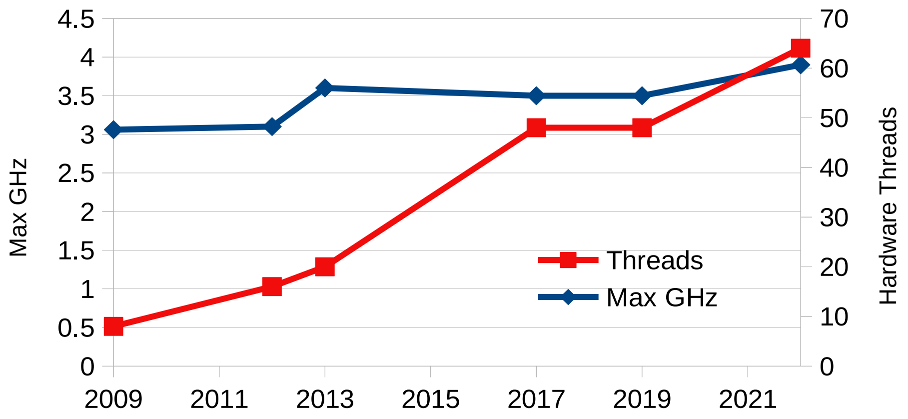
| Year | Processor | GHz |
|---|---|---|
| 1978 | Intel 8086 | 0.008 |
| 1985 | Intel 386 DX | 0.02 |
| 1993 | Intel Pentium | 0.06 |
| 1995 | Pentium Pro | 0.20 |
| 1999 | Pentium III | 0.50 |
| 2001 | Intel Xeon | 1.70 |



# Clock rate

Server Processor Examples (AWS EC2)

| Year | Processor | Cores/T. | Max GHz |
|---|---|---|---|
| 2009 | Xeon X5550 | 4/8 | 3.06 |
| 2012 | Xeon E5-2665 0 | 8/16 | 3.10 |
| 2013 | Xeon E5-2680 v2 | 10/20 | 3.60 |
| 2017 | Platinum 8175M | 24/48 | 3.50 |
| 2019 | Platinum 8259CL | 24/48 | 3.50 |
| 2022 | Xeon ? (R7iz*) | 32/64 | 3.90 |




## Increase has mostly leveled off due to power/efficiency

- (Blue line.) Workstation processors are higher; E.g., 2020 Xeon W-1270P @**5.1 GHz**

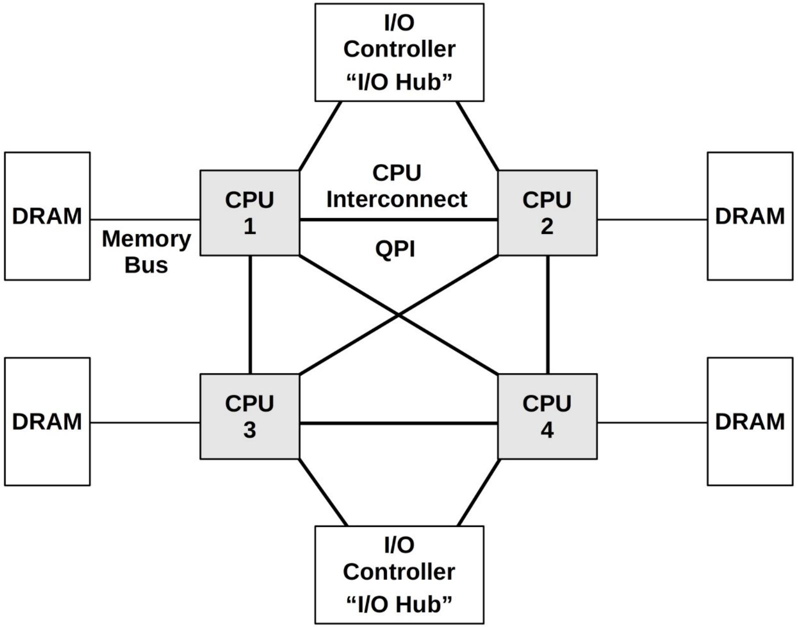
## Horizontal scaling instead

- (Red line.) More CPU cores, hardware threads, and server instances.

* R7iz launched one week ago and is still preview only; core/thread count is inferred [Barr 22]

# Interconnects

| Year | CPU Interconnect | Bandwidth Gbytes/s |
|---|---|---|
| 2007 | Intel FSB | 12.8 |
| 2008 | Intel QPI | 25.6 |
| 2017 | Intel UPI | 41.6 |




Source: Systems Performance 2nd Edition Figure 6.10 [Gregg 20]

## 10 years:

- 4x core count
- 3.25x bus rate

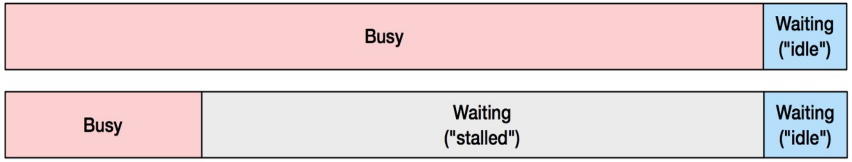
Memory bus (covered later) also lagging

## CPU utilization is wrong

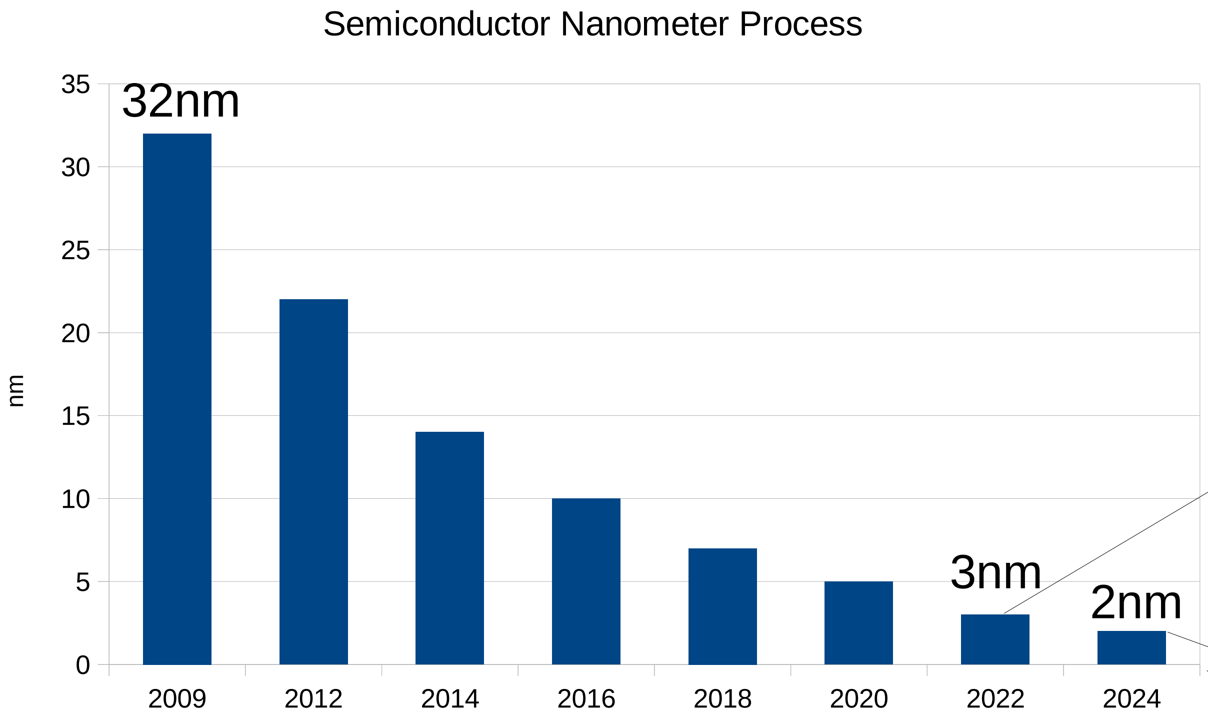
- Often mostly memory/interconnect stalls

90% CPU

...may mean:



# Lithography




Source: Semiconductor device fabrication [Wikipedia 21a]

TSMC expected volume production of 3nm in 2022 [Quach 21a], now expecting 2023 from Taiwan with a 4nm Arizona fab in 2024 [Gooding 22]

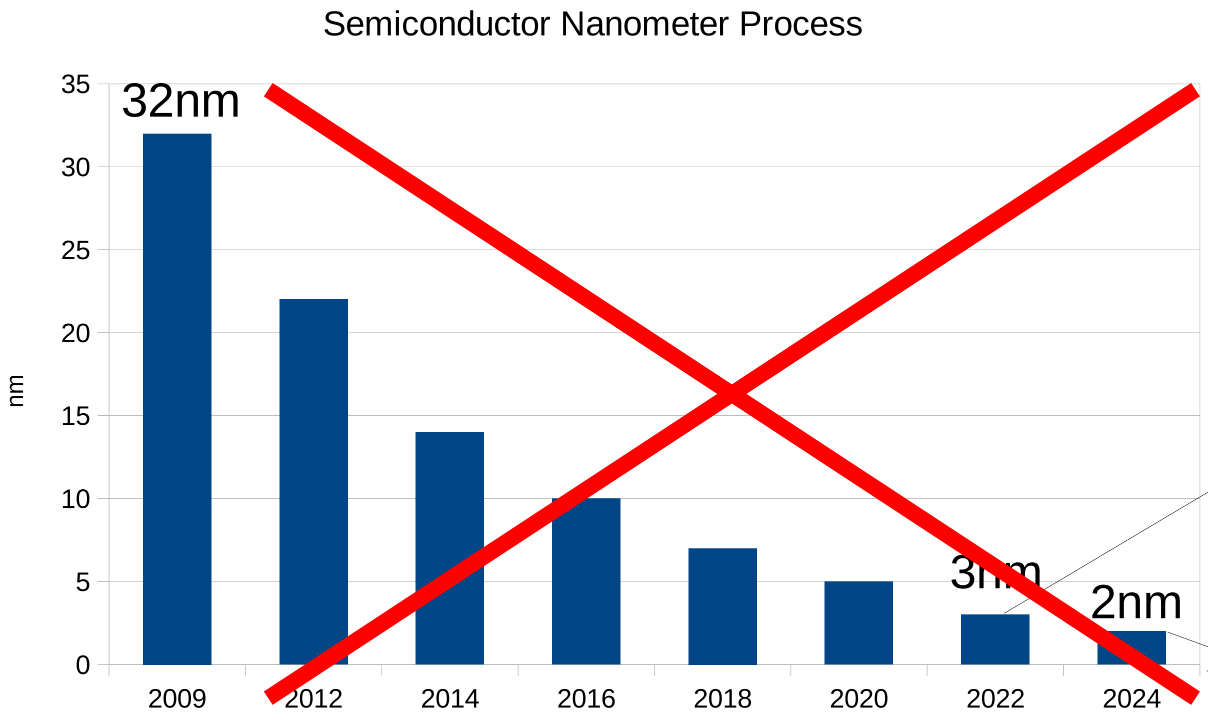
Meanwhile Intel building USD$20B Ohio "mega-fab" [Whalen 22]

IBM has already built one [Quach 21b]

Lithography limits expected to be reached by 2029, switching to stacked CPUs. [Moore 20]

BTW: Silicon atom diameter ~0.2 nm [Wikipedia 21b]

# Lithography




Source: Semiconductor device fabrication [Wikipedia 21a]

“Nanometer process” since 2010 should be considered a marketing term

New terms proposed include [Moore 20]:

- **GMT** (gate pitch, metal pitch, tiers)
- **LMC** (logic, memory, interconnects)

TSMC expected volume production of 3nm in 2022 [Quach 21a], now expecting 2023 from Taiwan with a 4nm Arizona fab in 2024 [Gooding 22]

Meanwhile Intel building USD$20B Ohio "mega-fab" [Whalen 22]

IBM has already built one [Quach 21b] (it has 12nm gate length)

Lithography limits expected to be reached by 2029, switching to stacked CPUs. [Moore 20]

BTW: Silicon atom diameter ~0.2 nm [Wikipedia 21b]

# Other processor scaling

## Special instructions

- E.g., AVX-512 Vector Neural Network Instructions (VNNI)

## Connected chiplets

- Using embedded multi-die interconnect bridge (EMIB) [Alcorn 17]. E.g., Intel Sapphire Rapids with 4 tiles [Tyson 21]; AMD Milan-X with 9 chiplets [Bonshor 22].

## 3D stacking

- E.g., Intel HBM, AMD Vcache [Cutress 21]

## Hybrid core architecture

- ARM big.LITTLE; Intel Alder Lake P-cores/E-cores [Alcorn 21]

# Recent server processor examples

| Vendor | Processor | Process | Clock | Cores/T. | LLC Mbytes | Date |
|---|---|---|---|---|---|---|
| Intel | Xeon Platinum 8380 (Ice Lake) | “10nm” | 2.3 - 3.4 | 40/80 | 60 | Apr 2021 |
| AMD | EPYC 9654P (Genoa) | “7nm” | 2.4 - 3.7 | 96/192 | 384 | Nov 2022 |
| ARM-based | Ampere Altra Max M128-30 | “7nm” | 3.0 | 128/128 | 32 | Sep 2021 |

Intel Alder Lake for server (Sapphire Rapids) coming soon. In preview on the Intel Developer Cloud [Intel 22] and AWS [Barr 22]. (Meanwhile: "Smuggler Hid Over 200 Alder Lake CPUs in Fake Silicone Belly" [Liu 22].)
Other server processors: IBM Z, RISC-V

## Coming soon to a datacenter near you

Although there is a **TSMC chip shortage** that may last through to 2022/2023 [Quatch 21][Ridley 21]

# Cloud chip race

## Amazon ARM/Graviton3

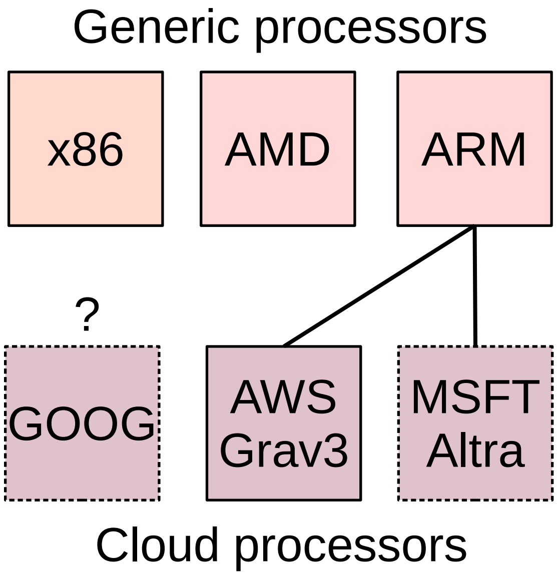
- ARM Neoverse V1, 64 core, 2.6 GHz
- **Graviton3E**: Customized for HPC

## Microsoft ARM/Ampere Altra

- ARM-based something was rumored [Warren 20]
- Ampere Altra-based types now launched in Azure [Nash 22]

## Google SoC

- Systems-on-Chip (SoC) coming soon [Vahdat 21]



# Accelerators

## GPUs

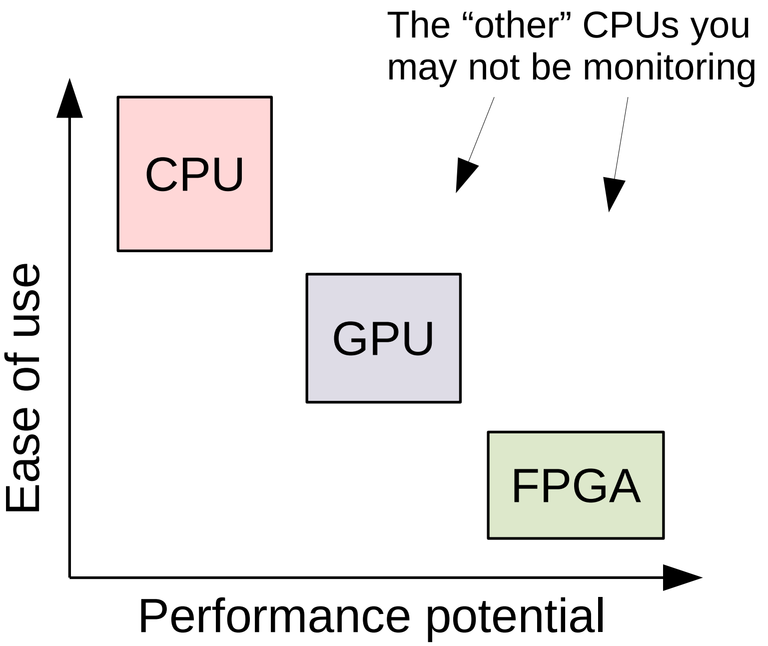
- Parallel workloads, thousands of GPU cores. Widespread adoption in machine learning.

## FPGAs

- Reprogrammable semiconductors
- Great potential, but needs specialists to program
- Good for algorithms: compression, cryptocurrency, video encoding, genomics, search, etc.
- Microsoft FPGA-based configurable cloud [Russinovich 17]

## Also IPUs, DPUs, TPUs, etc.

- Infrastructure processing units [Kummrow 21]
- Tensor processing units (TPU) [Google 21]
- AWS Trainium ML/AI accelerator Trn1n instances [Mann 22b]



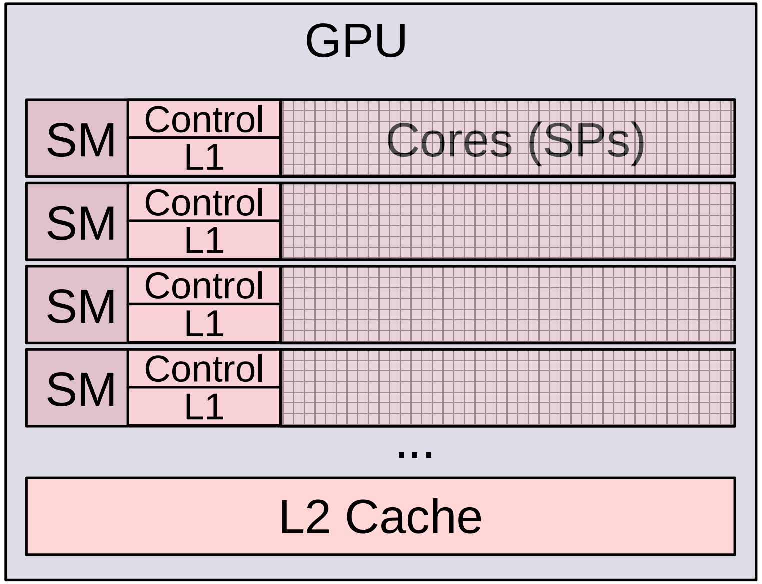
# Latest GPU examples

## NVIDIA GeForce RTX 3090: **10,496 CUDA cores**, 2020

- [Burnes 20]

## Cerebras Gen2 WSE: **850,000 AI-optimized cores**, 2021

- Use most of the silicon wafer for one chip.
  2.6 trillion transistors, 23 kW. [Trader 21]
- Previous version was already the “Largest chip ever built,” and US$2M. [insideHPC 20]
- Can now cluster them with Cerebras Wafer-Scale Cluster for millions of cores [Cerebras 22]




SM: Streaming multiprocessor

SP: Streaming processor

# Latest FPGA examples

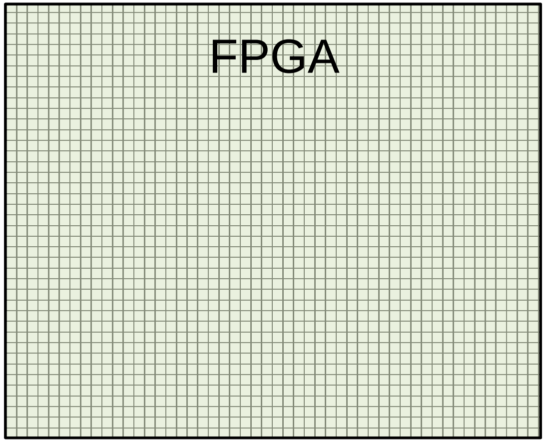
## Xilinx Virtex UltraScale+ VU19P, **8,938,000 logic cells**, 2019

- Using 35B transistors. Also has 4.5 Tbit/s transceiver bandwidth (bidir), and 1.5 Tbit/sec DDR4 bandwidth [Cutress 19]

## Xilinx Virtex UltraScale+ VU9P, **2,586,000 logic cells**, 2016

- Deploy right now: AWS EC2 F1 instance type (up to 8 of these FPGAs per instance)

AMD acquired Xilinx in 2022



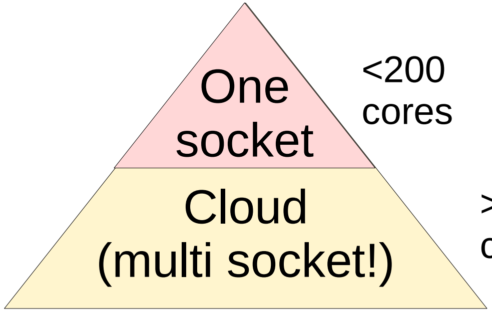

## BPF (covered later) already in FPGAs

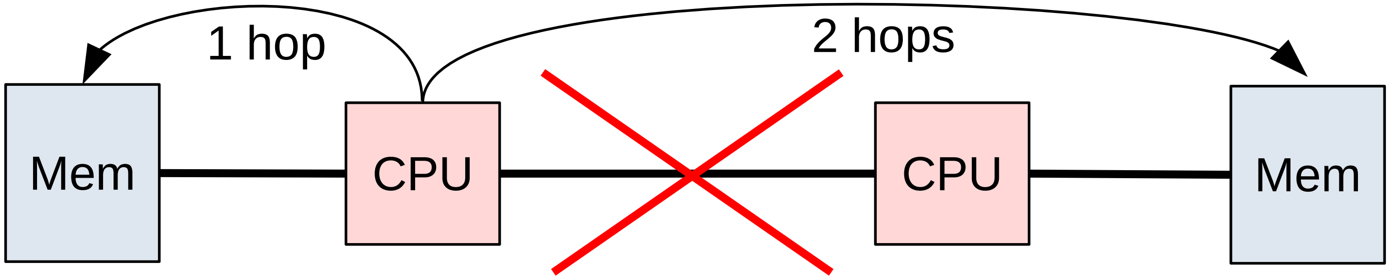
- E.g., 400 Gbit/s packet filter FFShark [Vega 20]

# My Predictions

# My Prediction: Multi-socket is doomed

- Single socket is getting big enough (cores)
- Already scaling horizontally (cloud)
  - And in datacenters, via “blades” or “microservers”
- Why pay NUMA costs?
  - Two 1-socket instances should out-perform one 2-socket instance
  - Multi-socket may hit some price/performance advantages given **rack/chassis overheads and costs**




Multi-socket future is mixed: one socket for cores, one GPU socket, one FPGA socket, etc. EMIB connected.

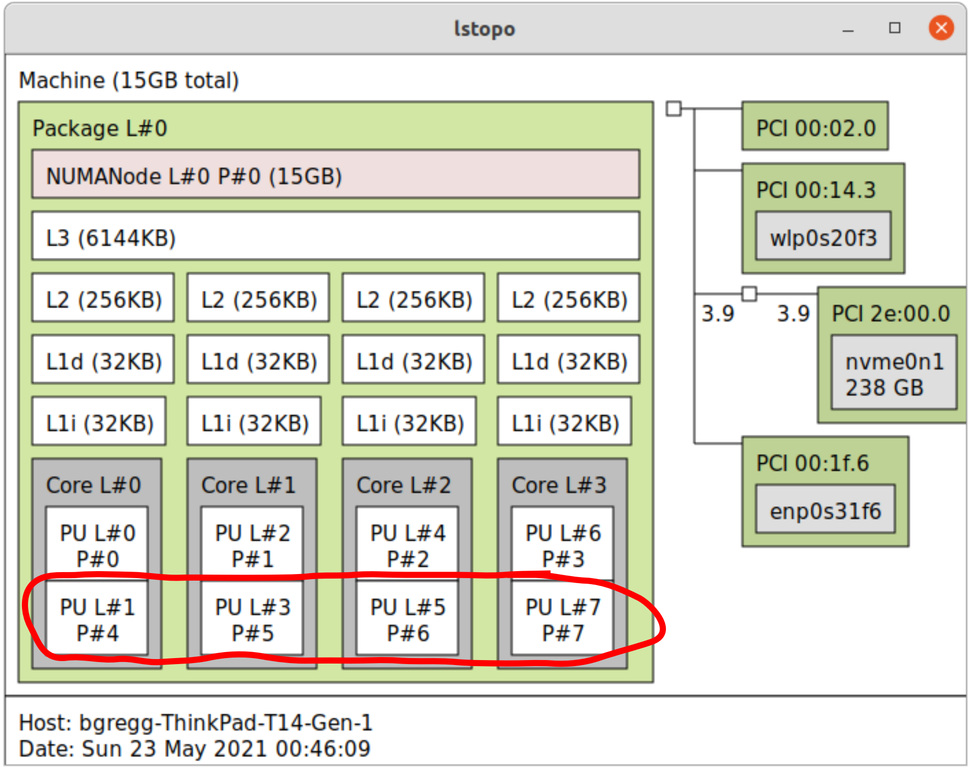
# My Prediction: SMT future unclear

Simultaneous multithreading (SMT) == hardware threads

- Performance variation
- ARM cores competitive
- Post meltdown/spectre
  - Some people turn them off

Possibilities:

- SMT becomes “free”
  - Processor feature, not a cost basis
  - Turn “oh no! hardware threads” into “great! bonus hardware threads!”
- No more hardware threads
  - Future investment elsewhere



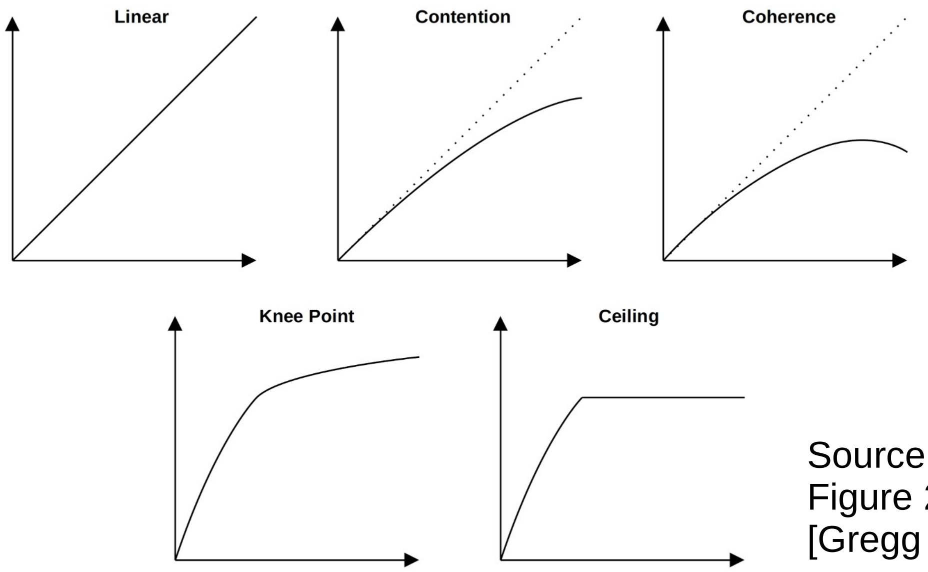
# My Prediction: Core count limits

*Imagine an 850,000-core server processor in today's systems...*

# My Prediction: Core count limits

Worsening problems:

- Memory-bound workloads
- Kernel/app lock contention
- False sharing
- Power consumption
- Core connectivity overheads
- etc.




Source:
Figure 2.16
[Gregg 20]

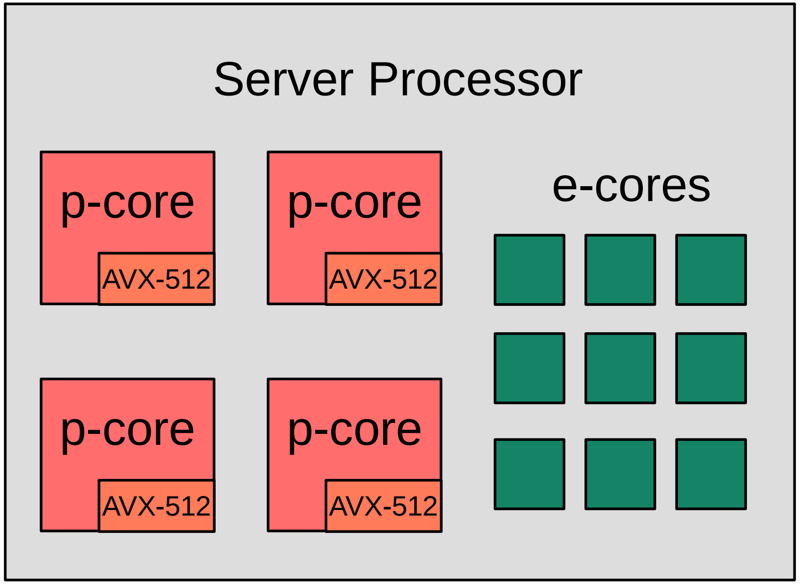
General-purpose computing will hit a **practical core limit**

- For a given memory subsystem & kernel, and running multiple applications
- E.g., 1024 cores (except GPUs/ML/AI); Esperanto RISC-V is already reaching “kilocore” scale [Kostovic 21]
- Apps themselves will hit an even smaller practical limit (some already have by design, e.g., Node.js and 2 CPUs)

# My Prediction: P-cores & E-cores

Intel Alder Lake (desktop) has performance and efficiency cores

This will come to server




Efficiency core tasks:

- Garbage collection
- NUMA rebalancing
- FS writeback compression & flushing
- Backups
- Security scanning
- etc.

Challenges include AVX-512

- Currently p-cores only, therefore cores aren't symmetric [Cutress 21b]. OS binary/scheduling challenges. Similar work: Linux 5.15 (2021) supports asymmetric scheduling for different ARM cores.

# My Prediction: 3 Eras of processor scaling

Delivered processor characteristics:

Era 1: Clock frequency

Era 2: Core/thread count

Era 3: Cache size & policy

# My Prediction: 3 Eras of processor scaling

Practical server limits:

Era 1: Clock frequency → **already reached by ~2005 (3.5 GHz)**

Era 2: Core/thread count → **limited by mid 2030s (e.g., 1024)**

Era 3: Cache size & policy → **limited by end of 2030s**

Mid-century will need an entirely new computer hardware architecture, kernel memory architecture, or logic gate technology, to progress further.

- E.g., use of graphine, carbon nanotubes [Hruska 12]
- This is after moving more to stacked processors

# My Prediction: More processor vendors

## ARM licensed or RISC-V

- Including Apple M1 for servers

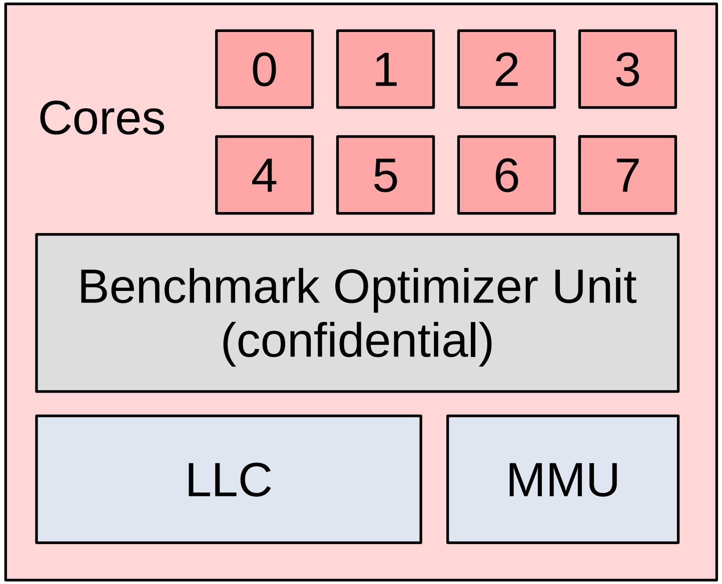
## Era of CPU choice

## Beware: “optimizing for the benchmark”

- Don’t believe microbenchmarks without doing “active benchmarking”: Root-cause perf analysis while the benchmark is still running.

## Intel making changes to compete

- Pat Gelsinger now CEO



DogeCPU “+AggressiveOpts” processor

# My Prediction: Cloud CPU advantage

## Large cloud vendors can analyze >100,000 workloads *directly*

- Via PMU PMCs and other processor features.

## Vast real-world detail to aid processor design

- More detail than traditional processor vendors have, and detail available immediately whenever they want.
- Will processor vendors offer their own clouds?
  - **Intel Developer Cloud** launched Sep 2022 for early access to chips and software [Robinson 22]

## Machine-learning aided processor design

- Based on the vast detail. Please point it at real-world workloads and not microbenchmarks.

Vast detail example: processor trace showing timestamped instructions:

```
# perf script --insn-trace --xed
date 31979 [003] 653971.670163672: ... (/lib/x86_64-linux-gnu/ld-2.27.so) mov %rsp, %rdi
date 31979 [003] 653971.670163672: ... (/lib/x86_64-linux-gnu/ld-2.27.so) callq 0x7f3bfbf4dea0
date 31979 [003] 653971.670163672: ... (/lib/x86_64-linux-gnu/ld-2.27.so) pushq %rbp
[...]
```

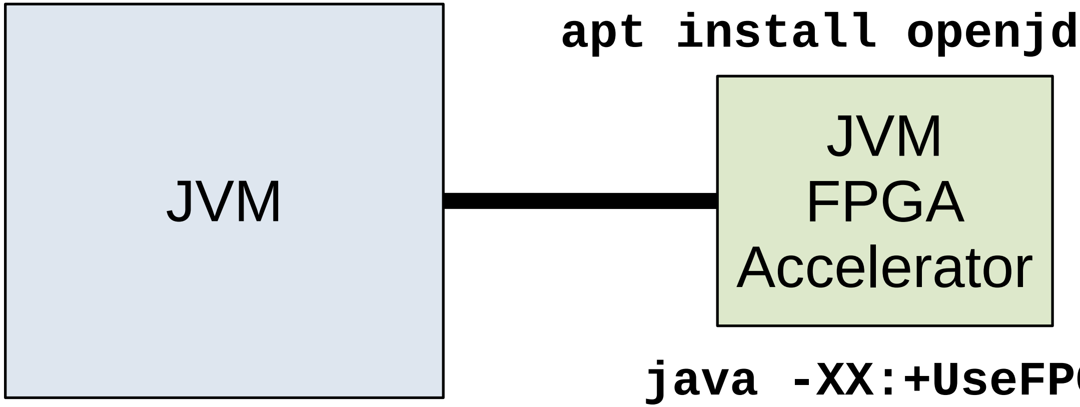
# My Prediction: FPGA turning point

## Little adoption (outside crypto & HFT) until major app support

- Solves the ease of use issue: Developers just configure the app (which may fetch and deploy an FMI)
- BPF use cases are welcome, but still specialized/narrow
- Needs runtime support, e.g., the JVM. Already work in this area (e.g., [TornadoVM 21]).

**`apt install openjdk-21`**

**`apt install openjdk-21-libfpga`**




**`java -XX:+UseFPGA`**

(none of this is real, yet)

# 2. Memory

# Many workloads memory I/O bound

```
# ./pmcarch 1
K_CYCLES   K_INSTR      IPC BR_RETIRED   BR_MISPRED  BMR% LLCREF      LLCMISS     LLC%
334937819  141680781   0.42 25744860335  536087729   2.08 1611987169  366692918  77.25
329721327  140928522   0.43 25760806599  525951093   2.04 1504594986  350931770  76.68
330388918  141393325   0.43 25821331202  484397356   1.88 1535130691  350629915  77.16
329889409  142876183   0.43 26506966225  510492279   1.93 1501785676  354458409  76.40
[...]
```

```
# ./pmcarch 1
K_CYCLES   K_INSTR      IPC BR_RETIRED   BR_MISPRED  BMR% LLCREF      LLCMISS     LLC%
38222881   25412094    0.66 4692322525   91505748    1.95 780435112   117058225  85.00
40754208   26308406    0.65 5286747667   95879771    1.81 751335355   123725560  83.53
35222264   24681830    0.70 4616980753   86190754    1.87 709841242   113254573  84.05
38176994   26317856    0.69 5055959631   92760370    1.83 787333902   119976728  84.76
[...]
```

```
# ./pmcarch
K_CYCLES   K_INSTR      IPC BR_RETIRED   BR_MISPRED  BMR% LLCREF      LLCMISS     LLC%
122697727  13892225    0.11 2604221808   40692664    1.56 419652590   93646793   77.68
144881903  17918325    0.12 3240599094   48088436    1.48 489936685   104672186  78.64
95561140   13815722    0.14 2722513072   42575763    1.56 401658252   94214458   76.54
99311699   15034220    0.15 2815805820   41802209    1.48 386979370   84139624   78.26
[...]
```

# DDR5 has better bandwidth

## DDR5 has a faster bus

- But not width

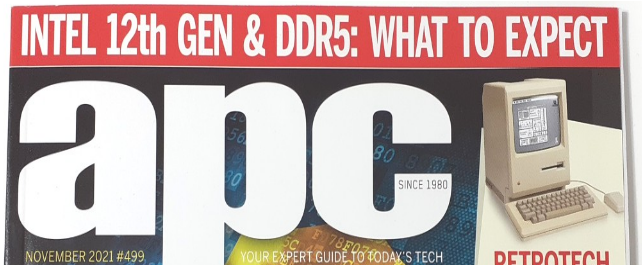
## 512GB DDR5 DIMMs

- Already released by Samsung [Shilov 21]

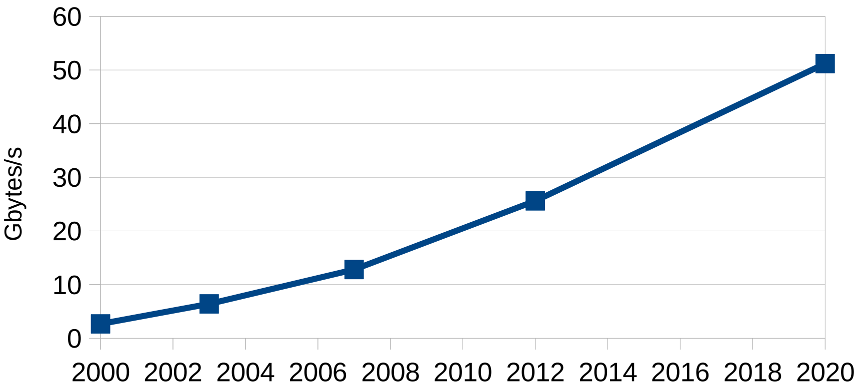
## Now arriving in clouds

- Needs processor support
- E.g., AWS Graviton2/3, Intel Sapphire Rapids

Desktop/Gamers have known for a while (Nov 2021):

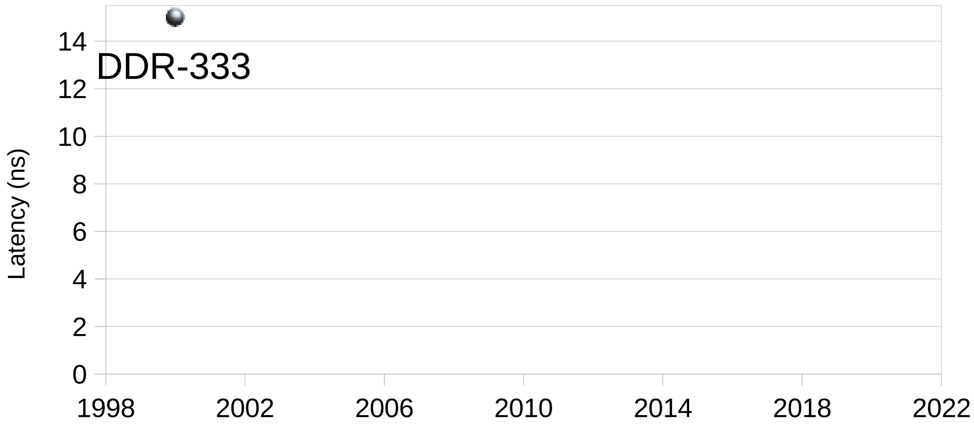



| Year | Memory | Peak Bandwidth Gbytes/s |
|---|---|---|
| 2000 | DDR-333 | 2.67 |
| 2003 | DDR2-800 | 6.4 |
| 2007 | DDR3-1600 | 12.8 |
| 2012 | DDR4-3200 | 25.6 |
| 2020 | DDR5-6400 | 51.2 |



# DDR latency

| Year | Memory | Latency (ns) |
|---|---|---|
| 2000 | DDR-333 | 15 |
| | | |
| | | |
| | | |
| | | |



# DDR latency

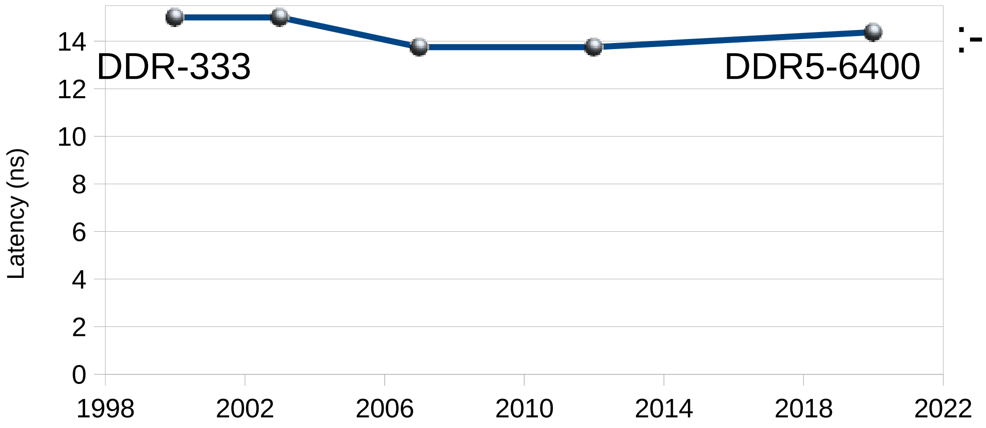
## Hasn’t changed in 20 years

- This is single access latency
- Same memory clock (200 MHz) [Greenberg 11]
- Also see [Cutress 20][Goering 11]

## Low-latency DDR does exist

- Reduced Latency DRAM (RLDRAM) by Infineon and Micron: lower latency but lower density
- Not seeing widespread server use (I’ve seen it marketed towards HFT)

| Year | Memory | Latency (ns) |
|---|---|---|
| 2000 | DDR-333 | 15 |
| 2003 | DDR2-800 | 15 |
| 2007 | DDR3-1600 | 13.75 |
| 2012 | DDR4-3200 | 13.75 |
| 2020 | DDR5-6400 | 14.38 |




:-(

# HBM

## High bandwidth memory, 3D stacking

- Target uses cases include high performance computing, and virtual reality graphical processing [Macri 15]

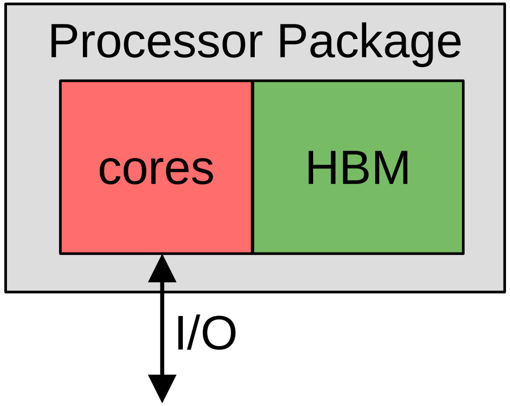
## GPUs already use it

## Processors now including HBM on-package

- **Intel Sapphire Rapids** (Xeon Max) has 64 Gbytes of HBM2e in 4 clusters, for >1 Tbyte/s memory bandwidth and >1 Gbyte per core [Pirzada 22]

## No DRAM systems now possible!

- Intel's 3 modes:
  - HBM Only: No DRAM
  - HBM Flat: 2 memory regions, software to optimize placement
  - HBM Caching: HBM caches DDR

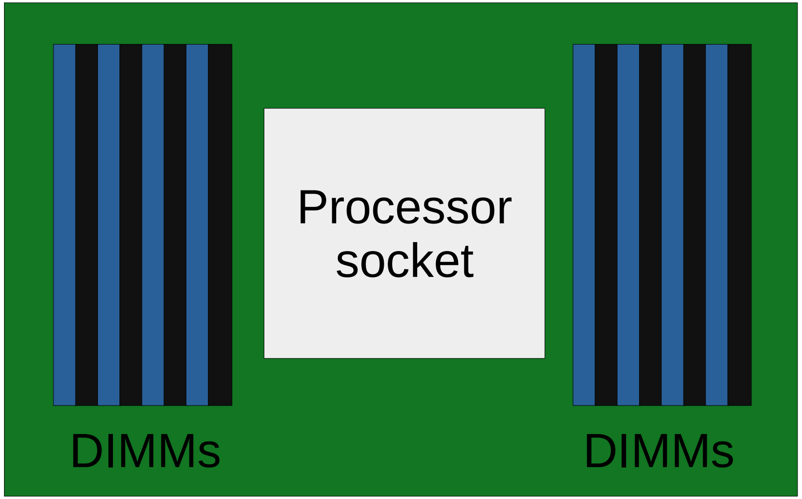


# Server DRAM size

## SuperMicro SuperServer B12SPE-CPU-25G

- Single Socket (see earlier slides)
- 16 DIMM slots
- **4 TB DDR-4**

[SuperMicro 21]




B12SPE-CPU-25G

## Facebook Delta Lake (1S) OCP

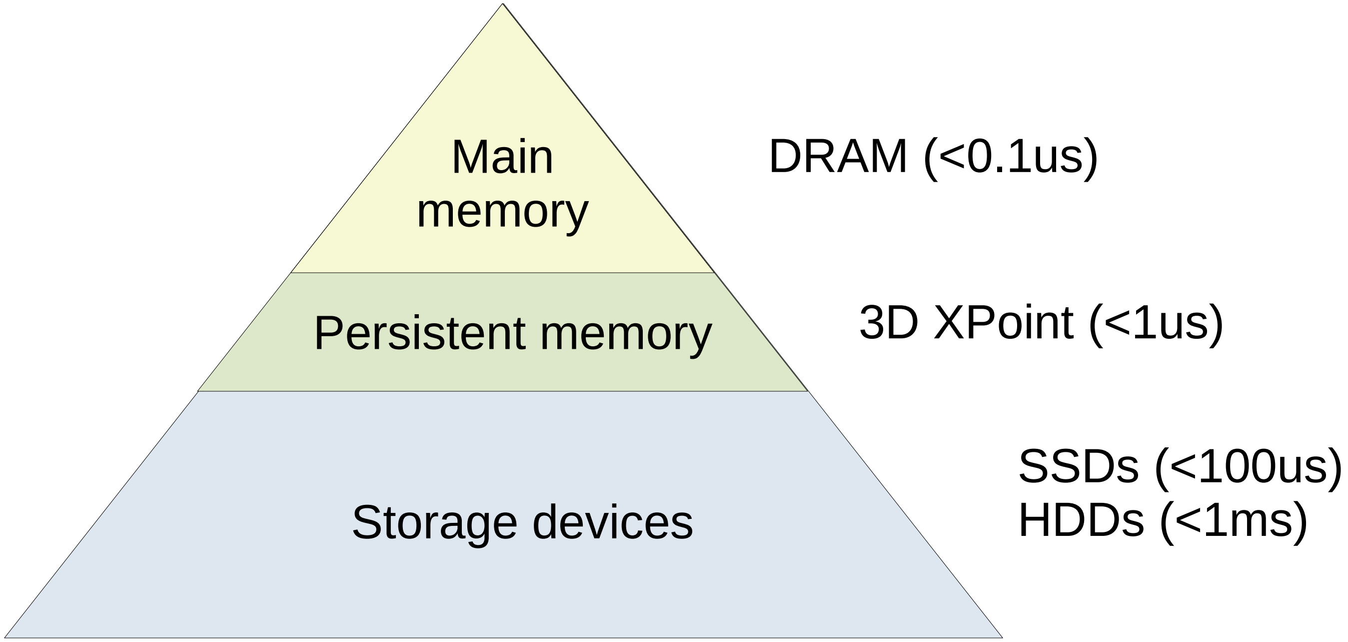
- 6 DIMM slots
- **96 Gbytes DDR-4**
- Price/optimal for a typical WSS?

[Haken 21]

# Additional memory tier not successful

## Intel/Micron's 3D XPoint now cancelled [Mann 22]

- Could also operate in application direct mode and storage mode [Intel 21]



# My Prediction: Extrapolation

Not a JEDEC announcement

Assumes miraculous engineering work

- For various challenges see [Peterson 20]

But will single-access latency drop in DDR-6?

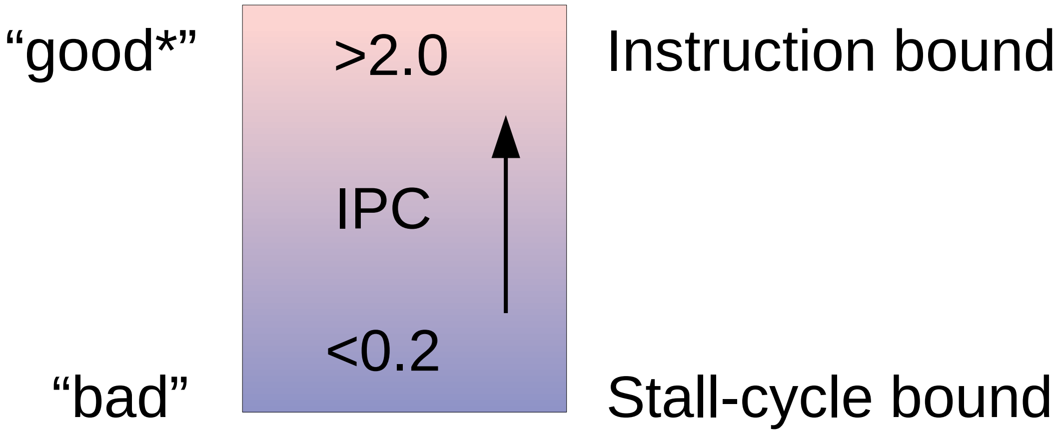
- I'd guess not, DDR internals are already at their cost-sweet-spot, leaving low-latency for other memory technologies

| Year | Memory | Peak Bandwidth Gbytes/s |
|---|---|---|
| 2000 | DDR-333 | 2.67 |
| 2003 | DDR2-800 | 6.4 |
| 2007 | DDR3-1600 | 12.8 |
| 2012 | DDR4-3200 | 25.6 |
| 2020 | DDR5-6400 | 51.2 |
| **2028** | **DDR6-12800** | **102.4** |
| **2036** | **DDR7-25600** | **204.8** |
| **2044** | **DDR8-51200** | **409.6** |

doubling ↓

# My Prediction: DDR5 “up to 2x” Wins

E.g., IPC 0.1 → ~0.2 for *bandwidth*-bound workloads




* probably; exceptions include spin locks

If DDR-6 gets a latency drop, more frequent wins

# My Prediction: HBM-only servers

## Clouds offering “high bandwidth memory” HBM-only instances

- HBM on-processor
- Finally helping memory catch up to core scaling

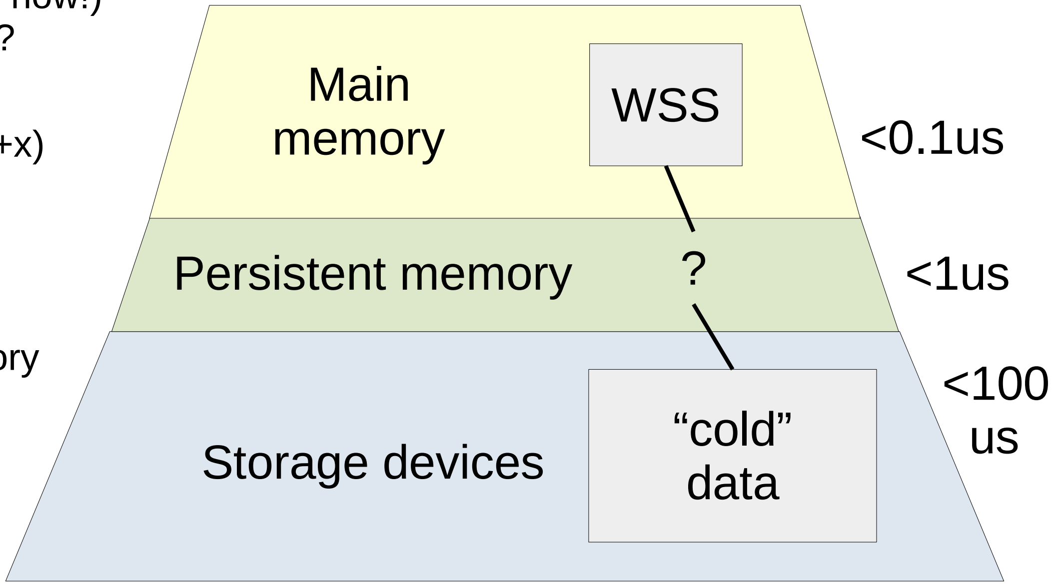
## RLDRAM on-package as another option?

- “Low latency memory” instance

# My *Prior* Prediction: Extra tier too late

## Competition isn't disks, it's Tbytes of DRAM

- SuperMicro's single socket should hit 8 Tbytes DDR-5
- AWS EC2 p4.24xl has 1.1 Tbytes of DRAM (deploy now!)
  How often does your working set size (WSS) not fit?
  Across several of these for redundancy?
- Next tier needs to get much bigger than DRAM (10+x) and much cheaper to find an extra-tier use case (e.g., cost based).
- Meanwhile, DRAM is still getting bigger and faster
- I developed the first cache tier between main memory and disks to see widespread use:
  the ZFS L2ARC [Gregg 08]




## It's more like a trapezoid

# 3. Disks

# Recent timeline for rotational disks

## 2005: Perpendicular magnetic recording (PMR)

- Writes vertically using a shaped magnetic field for higher density

## 2013: Shingled magnetic recording (SMR)

- (next slide)

## 2019: Multi-actuator technology (MAT)

- Two sets of heads and actuators; like 2-drive RAID 0 [Alcorn 17].

## 2020: Energy-assisted magnetic recording (EAMR)

- Western Digital 18TB & 20TB [Salter 20]

## 2021: Heat-assisted magnetic recording (HAMR)

- Seagate 20TB HAMR drives [Shilov 21b]

# Recent timeline for rotational disks

## 2005: Perpendicular magnetic recording (PMR)

- Writes vertically using a shaped magnetic field for higher density

## 2013: Shingled magnetic recording (SMR)

- (next slide)

## 2019: Multi-actuator technology (MAT)

- Two sets of heads and actuators; like 2-drive RAID 0 [Alcorn 17].

## 2020: Energy-assisted magnetic recording (EAMR)

- Western Digital 18TB & 20TB [Salter 20]

## 2021: Heat-assisted magnetic recording (HAMR)

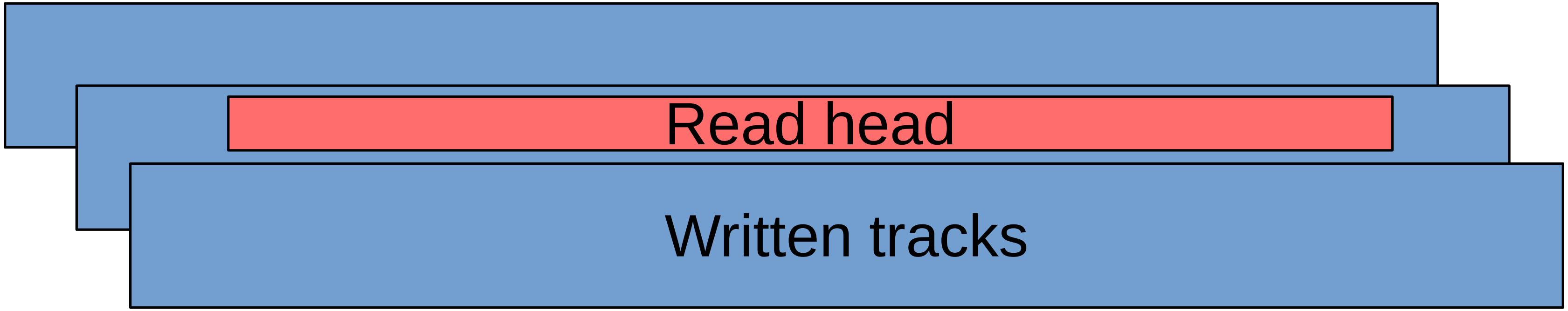
- Seagate 20TB HAMR drives [Shilov 21b]

I don’t know their perf characteristics yet

# SMR

## 11-25% more storage, worse performance

- Writes tracks in an overlapping way, like shingles on a roof. [Shimpi 13]
- Overwritten data must be rewritten. Suited for archival (write once) workloads.




## Look out for 18TB/20TB-with-SMR drive releases

# Flash memory-based disks

Single-Level Cell (SLC)

Multi-Level Cell (MLC)

Enterprise MLC (eMLC)

2009: Tri-Level Cell (TLC)

2009: Quad-Level Cell (QLC)

- QLC is only rated for around 1,000 block-erase cycles [Liu 20].

2013: 3D NAND / Vertical NAND (V-NAND)

- SK Hynix envisions 600-Layer 3D NAND [Shilov 21c]. Should be multi-Tbyte.

**2017-2022**: Intel Optane (3D XPoint persistent memory) disks, now cancelled; was used as an accelerator

SSD performance pathologies: latency from aging, wear-leveling, fragmentation, internal compression, etc.

# Storage Interconnects

## SAS-4 cards in development

- (Storage attached SCSI)

## PCIe 5.0 coming soon

- (Peripheral Component Interconnect Express)
- Intel already demoed on Sapphire Rapids [Hruska 20]

## NVMe 1.4 latest

- (Non-Volatile Memory Express)
- Storage over PCIe bus
- Support zoned namespace SSDs (ZNS) [ZonedStorage 21]
- Bandwidth bounded by PCIe bus

## These have features other than speed

- Reliability, power management, virtualization support, etc.

| Year Specified | Interface | Bandwidth Gbit/s |
|---|---|---|
| 2003 | SAS-1 | 3 |
| 2009 | SAS-2 | 6 |
| 2012 | SAS-3 | 12 |
| 2017 | SAS-4 | 22.5 |
| **202?** | **SAS-5** | **45** |

| Year Specified | Interface | Bandwidth 16 lane Gbyte/s |
|---|---|---|
| 2003 | PCIe 1 | 4 |
| 2007 | PCIe 2 | 8 |
| 2010 | PCIe 3 | 16 |
| 2017 | PCIe 4 | 31.5 |
| **2019** | **PCIe 5** | **63** |

# Latest storage device examples

## 2022 SSD: Samsung PM1743 [Smith 22]

- Up to 15.36 Tbytes
- PCIe Gen5
- Sequential reads up to 13 Gbytes/sec

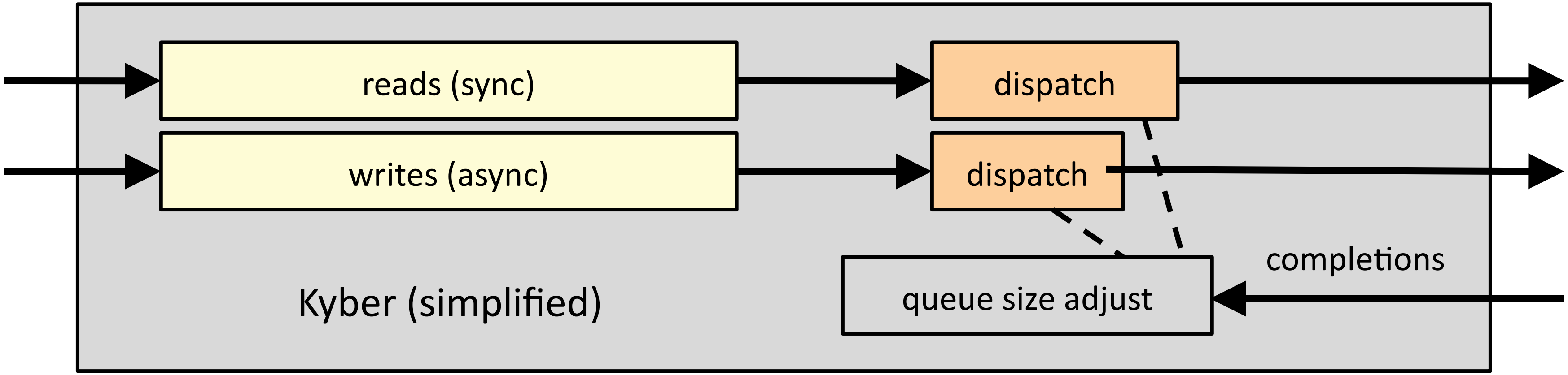
## 2022 HDD: Seagate Exos 2x18 [Seagate 22]

- 18 or 16 Tbytes
- Helium sealed
- Multi-actuator (2 x sets of heads) for "2x more performance"
- Up to 554 Mbytes/sec: "SSD performance"
- 7200 RPM

# Linux Kyber I/O scheduler

## Multi-queue, target read & write latency

- Up to 300x lower 99th percentile latencies [Gregg 18]
- Linux 4.12 [Corbet 17]



# My Prediction: Slower rotational

## Archive focus

- There’s ever-increasing demand for storage (incl. social video today; social VR tomorrow?)
- Needed for archives
- More “weird” pathologies. SMR is just the start.
- Even less tolerant to shouting

Bigger, slower, and weirder

# My Prediction: More flash pathologies

- Worse internal lifetime
- More wear-leveling & logic
- More latency outliers

Bigger, *faster*, and weirder

We need more observability of flash drive internals

# 4. Networking

# Latest Hardware

## 400 Gbit/s in use

- E.g., 400 Gbit/s switches/routers by Cisco and Juniper, tranceivers by Arista and Intel
- AWS EC2 P4 instance type (deploy now!)
- On PCI, needs PCIe 5

## 800 Gbit/s next

- [Charlene 20]
- Terabit Ethernet (1 Tbit/s) not far away

## More NIC features

- E.g., inline kTLS (TLS offload to the NIC), e.g., Mellanox ConnectX-6-Dx [Gallatin 19]
- **FPGA, P4, and eBPF support.**

# Protocols

## QUIC / HTTP/3

- TCP-like sessions over (fast) UDP.
- 0-RTT connection handshakes. For clients that have previously communicated.

## MP-TCP

- Multipath TCP. Use multiple paths in parallel to improve throughput and reliability. RFC-8684 [Ford 20]
- Linux support starting in 5.6.

# Linux TCP Congestion Control Algorithms

## DCTCP

- Data Center TCP. Linux 3.18. [Borkmann 14]

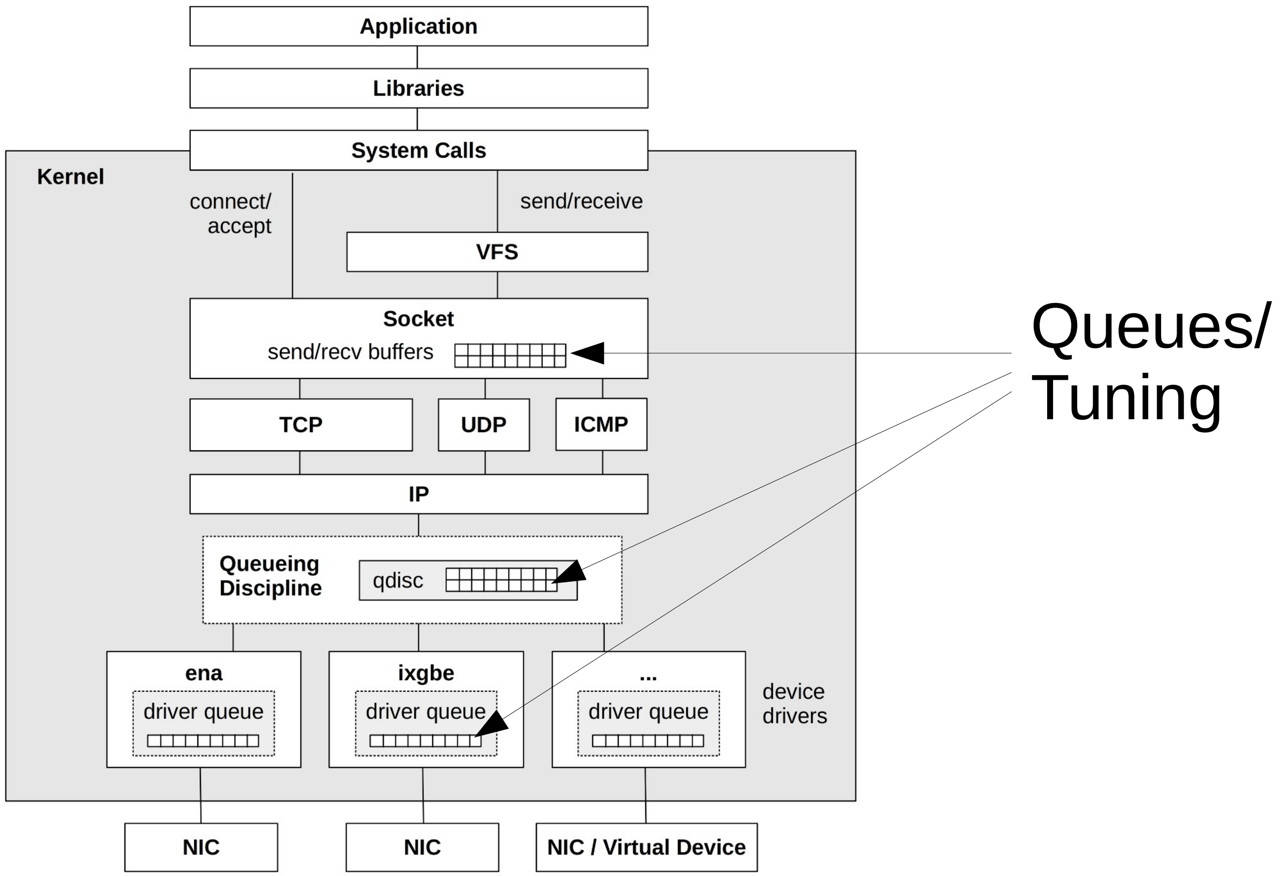
## TCP NV

- New Vegas. Linux 4.8

## TCP BBR

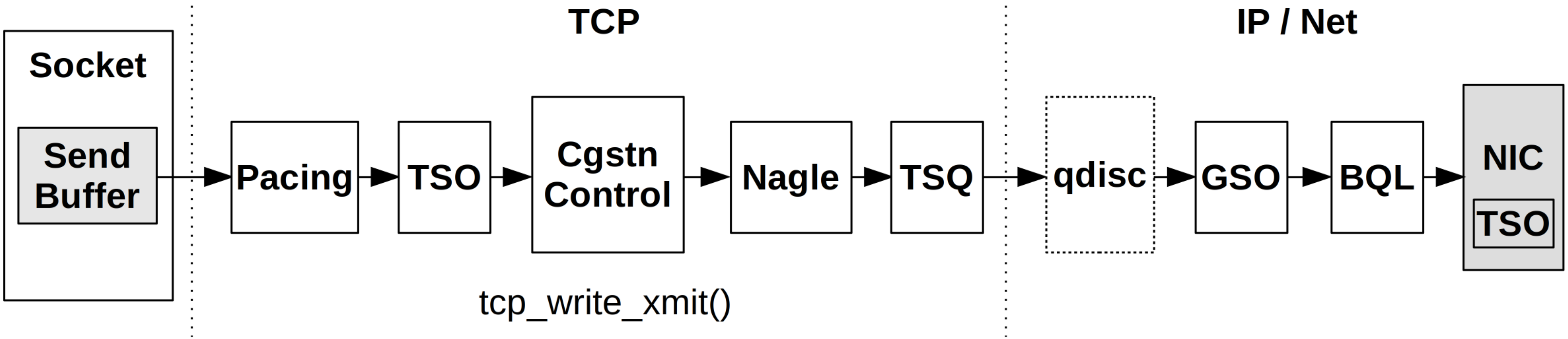
- Bottleneck Bandwidth and RTT (BBR) improves performance on packet loss networks [Cardwell 16]
- With 1% packet loss, Netflix sees 3x better throughput [Gregg 18]

# Linux Network Stack




Source: Systems Performance 2nd Edition, Figure 10.8 [Gregg 20]

# Linux TCP send path




Keeps adding performance features

Source: Systems Performance 2nd Edition, Figure 10.11 [Gregg 20]

# Software

## eXpress Data Path (XDP) (uses eBPF)

- Programmable fast lane for networking. In the Linux kernel.
- A role previously served by DPDK and kernel bypass.

# My Prediction: BPF in FPGAs/IPUs

Massive I/O tranceiver capabilities

Netronome already did BPF in hardware

Edge computing on the NICs

# My Prediction: Cheap BPF routers

## Linux + BPF + 400 GbE NIC

- Cheap == commodity hardware
- Use case from the beginning of eBPF (PLUMgrid)

# My Prediction: More demand for network perf

Apps increasingly network

World of sensors

Remote work & video conferencing

Netflix 4K content

VR tourism & multiverse

# 5. Kernels

# Latest Kernels/OSes

May 2022: FreeBSD 13.1

Oct 2022: Linux 6.0 ("Hurr durr I'ma ninja sloth" [Torvalds 22])

Nov 2022: Windows 22H2 (10.0.22621.900)

# Recent Linux perf features

2022: IPv6 jumbograms, packets >64 Kbytes (5.19)

2021: BPF kernel function calls, e.g., for TCP cong ctrl (5.13)

2020: Static calls to improve Spectre-fix (5.10)

2020: BPF on socket lookups (5.9)

2020: Thermal pressure (5.7)

2020: MultiPath TCP (5.6)

2019: MADV_COLD, MADV_PAGEOUT (5.4)

2019: io_uring (5.1)

2019: UDP GRO (5.0)

2019: Multi-queue I/O default (5.0)

2018: TCP EDT (4.20)

2018: PSI (4.20)

Plus lots more, including support for the latest x86/AMD/ARM/etc. instructions (e.g., AMX in 5.16, LoongArch in 5.19)

For 2016-2018, see my summary: [Gregg 18].
Includes CPU schedulers (thermal, topology);
Block I/O qdiscs; Kyber scheduler (earlier slide);
TCP congestion control algoritms (earlier slide); etc.

# Recent Linux perf features

2022: IPv6 jumbograms, packets >64 Kbytes (5.19)

2021: BPF kernel function calls, e.g., for TCP cong ctrl (5.13)

2020: Static calls to improve Spectre-fix (5.10)

2020: BPF on socket lookups (5.9)

2020: Thermal pressure (5.7)

2020: MultiPath TCP (5.6)

2019: MADV_COLD, MADV_PAGEOUT (5.4)

2019: io_uring (5.1)

2019: UDP GRO (5.0)

2019: Multi-queue I/O default (5.0)

2018: TCP EDT (4.20)

2018: PSI (4.20)

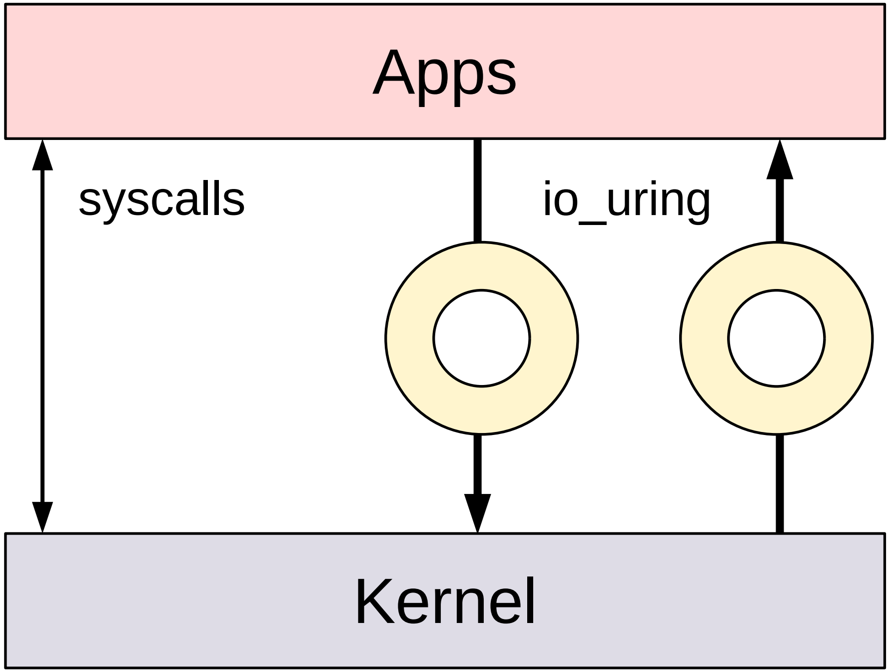
Plus lots more, including support for the latest x86/AMD/ARM/etc. instructions (e.g., AMX in 5.16, LoongArch in 5.19)

For 2016-2018, see my summary: [Gregg 18]. Includes CPU schedulers (thermal, topology); Block I/O qdiscs; Kyber scheduler (earlier slide); TCP congestion control algoritms (earlier slide); etc.

# io_uring

## Faster syscalls using shared ring buffers

- Send and completion ring buffers
- Allows I/O to be batched and async
- Primary use cases network and disk I/O



# eBPF Everywhere


Microsoft Microsoft 365 Azure Office 365 Dynamics 365 Power Platform Windows 10

Microsoft Open Source Blog

## Making eBPF work on Windows

May 10, 2021

Share

Dave Thaler

Partner Software Engineer, Microsoft

Poorna Gaddehosur

Principal Software Engineer Lead, Microsoft

eBPF is a well-known but revolutionary technology—providing programmability,

[Thaler 21]

Plus eBPF for BSD projects already started.

# eBPF == BPF

2015:

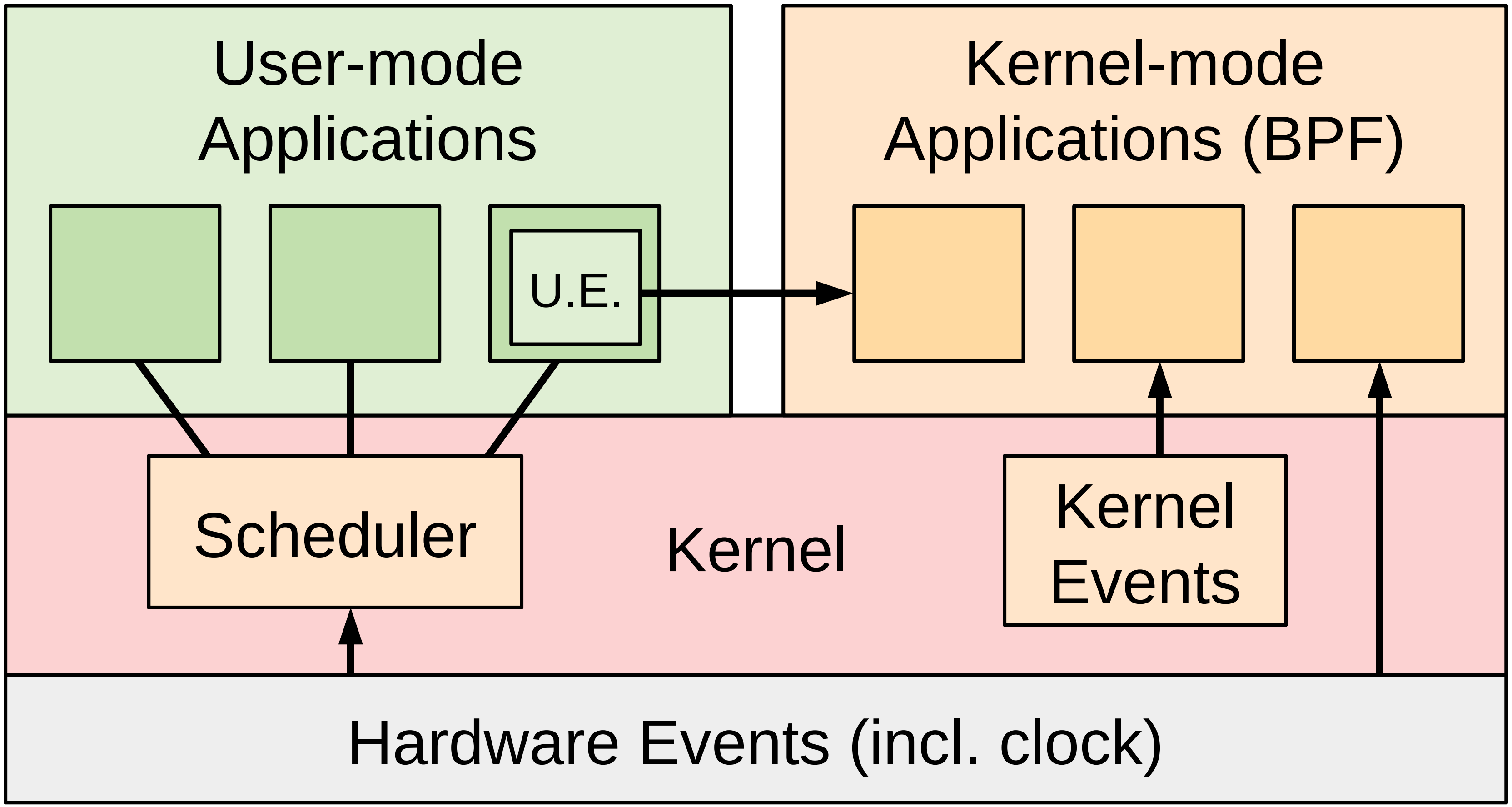
- BPF: Berkeley Packet Filter
- eBPF: extended BPF

2022:

- “Classic BPF”: Berkeley Packet Filter
- BPF: A technology name (aka eBPF)
  - Kernel engineers like to use “BPF”; companies “eBPF”.

This is what happens when you don’t have marketing professionals help name your product.

# BPF Future: Event-based Applications




https://twitter.com/srostedt/status/1177147373283418112

# Emerging BPF uses

Observability agents

Security intrusion detection, zero-day mitigation

TCP congestion control algorithms

Application accelerators

- E.g., Orange bmc-cached accelerator for memcached [Ghigoff 21]

General kernel development

# My Prediction: Future BPF Uses

File system buffering/readahead policies

CPU scheduler policies

Lightweight I/O-bound applications (e.g., proxies)

- Or such apps can go to io_uring or FPGAs. “Three buses arrived at once.”
  - When I did engineering at University: “people ride buses and electrons ride busses.” Unfortunately that usage has gone out of fashion, otherwise it would have been clear which bus I was referring to!

# My Prediction: Easy/Auto PGO

## PGO/AutoFDO shows ~10% wins, but hard to manage

- Performance-guided optimization (PGO) / Auto feedback-directed optimization (AutoFDO)
- Some companies already do kernel PGO (Google [Tolvanen 20], Microsoft [Bearman 20])
- We can't leave 10% on the table forever, someone will do an easy-PGO product or it becomes a feature of AI auto-tuners.

## Partial JIT support?

# My Prediction: Kernel emulation often slow

## I can run <kernel> apps under <other kernel>

## by emulating <a bare-minimal set of> syscalls!

Cool project, but:

- Missing latest kernel and perf features (E.g., Linux’s BPF, io_uring, WireGuard, etc. Plus certain syscall flags return ENOTSUP. So it’s like a weird old fork of Linux.)
  - Some exceptions: E.g., another kernel may have better hardware support, which may benefit apps more than the loss of kernel capabilities.
- Debugging and security challenges. Better ROI with lightweight VMs.

In other words, WSL2 >> WSL1

# My Prediction: OS performance

## Linux: increasing complexity & worse perf defaults

- Linux at FAANGs and other large companies is often very different to the Linux publicly available, as they have perf and OS teams who can configure advanced technologies and tune and fix things (like enabling frame pointers). This means most experts are not tuning the Linux *you* are using and various defaults get little attention and rot (e.g., high-speed network engineers configure XDP and QUIC, and aren't looking at defaults with TCP). A bit more room for a lightweight kernel (e.g., BSD) with better perf defaults to compete. Similarities: Oracle DB vs MySQL; MULTICS vs UNIX.

## BSD: high perf for narrow uses

- Still serving some companies (including Netflix) very well thanks to tuned performance (see footnote on p124 of [Gregg 20]). Path to growth is better EC2/Azure performance support, but it may take years before a big customer (with a perf team) migrates and gets everything fixed. There are over a dozen of perf engineers working on Linux on EC2; BSD needs at least one *full time* senior EC2 (not metal) perf engineer.

## Windows: community perf improvements

- BPF tracing support allows outsiders to root cause kernel problems like never before (beyond ETW/Xperf). Will have a wave of finding "low hanging fruit" to begin with, improving perf and reliability.

# My Prediction: Unikernels

## Finally gets *one* compelling published use case

“2x perf for X”

## But few people run X

- Needs to be really kernel heavy, and not many workloads are. And there’s already a lot of competition for reducing kernel overhead (BPF, io_uring, FPGAs, DPDK, etc.)
- Once one use case is found, it may form a valuable community around X and Unikernels. But it needs the published use case to start, preferably from a FAANG.
- Does need to be 2x or more, not 20%, to overcome the cost of retooling everything, redoing all observability metrics, profilers, etc. It’s not impossible, but not easy [Gregg 16].
- More OS-research-style wins found from hybrid- and micro-kernels.

# 6. Hypervisors

# Containers

## Cgroup v2 rollout

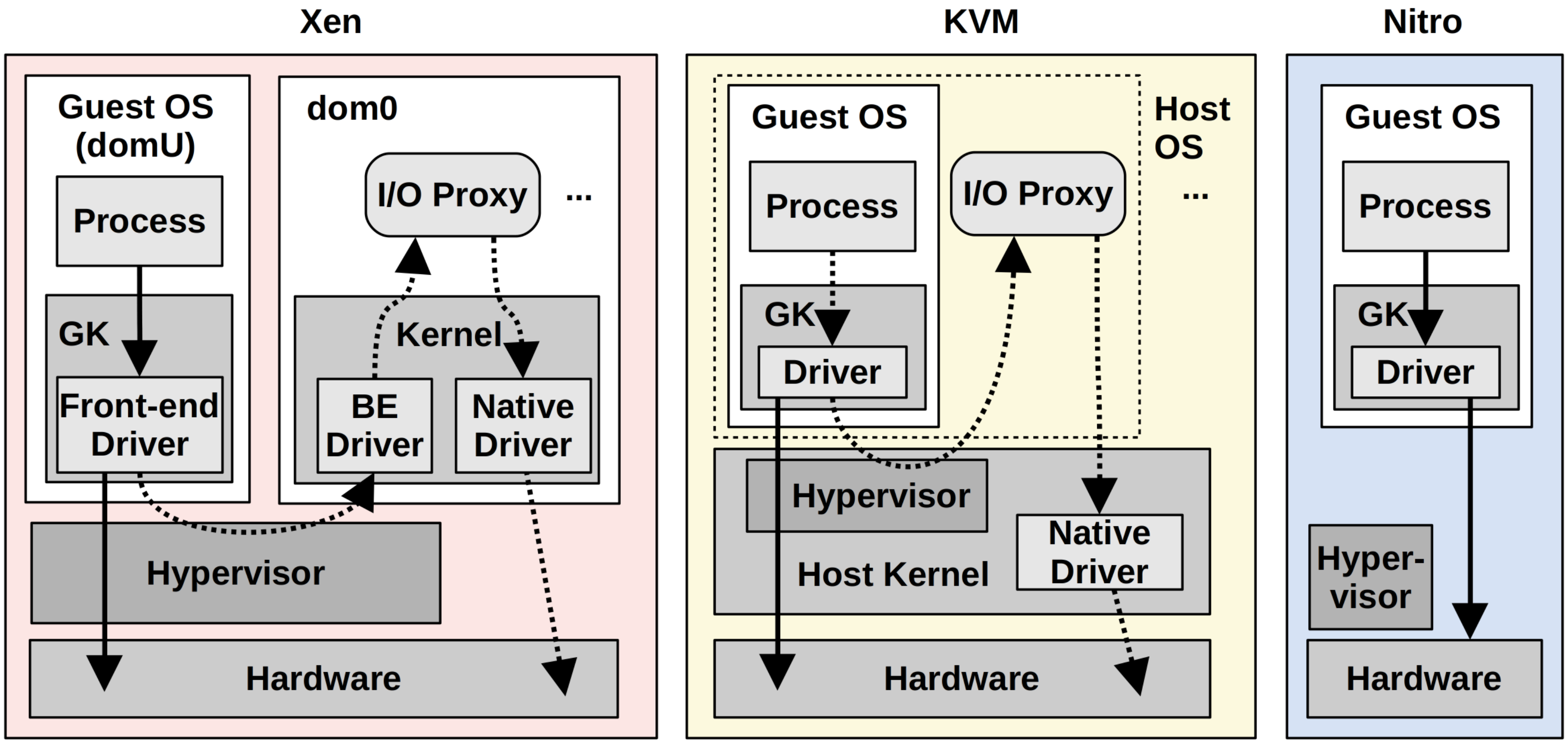
## Container scheduler adoption

- Kubernetes, OpenStack, and more
- Netflix develops its own called "Titus" [Joshi 18]
- Price/performance gains: "Tetris packing" workloads without too much interference (clever scheduler)

## Many perf tools still not "container aware"

- Usage in a container not restricted to the container, or not permitted by default (needs CAP_PERFMON CAP_SYS_PTRACE, CAP_SYS_ADMIN)

# Hardware Hypervisors




Source: Systems Performance 2nd Edition, Figure 11.17 [Gregg 20]

# VM Improvements

Bare-metal performance
Near-metal performance
Optimized performance
Poor performance

Importance: Most → Least

Old → New

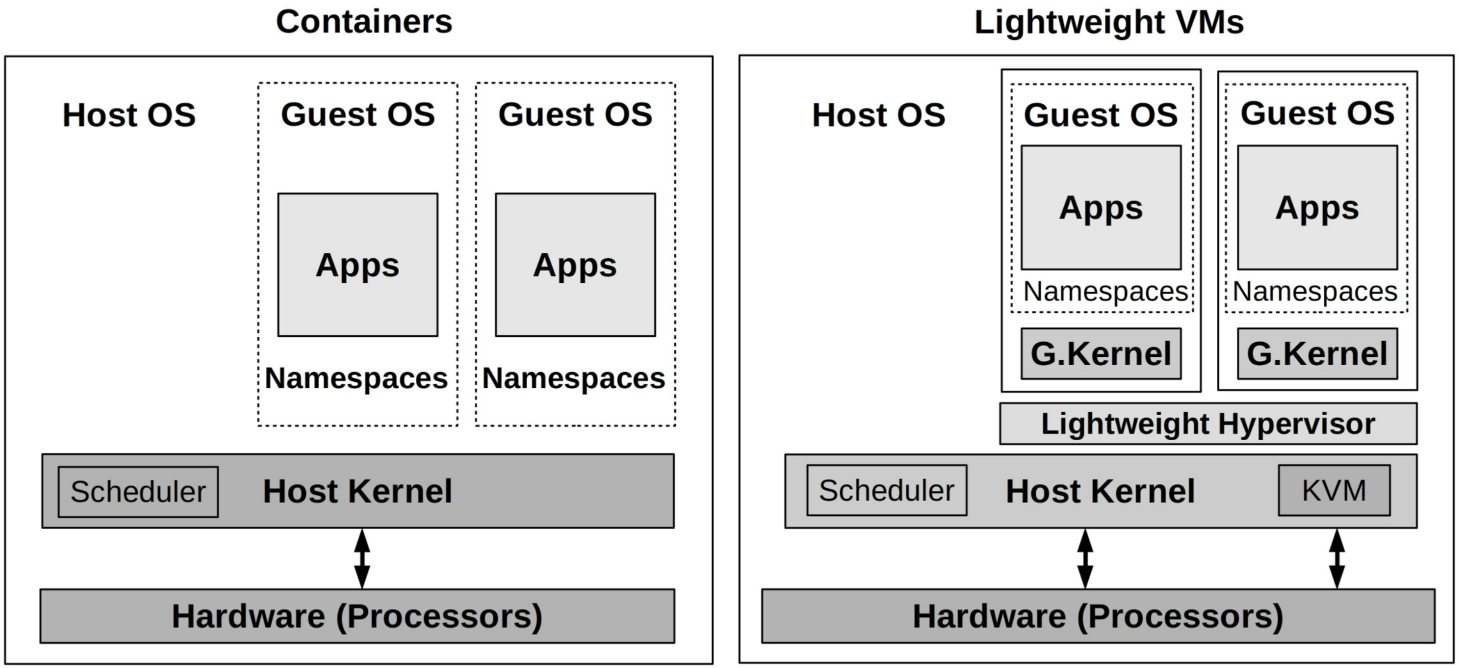
| # | Tech | Type | With | CPU, Memory | Network I/O | Local Storage I/O | Remote Storage I/O | Interrupts, Timers | Motherboard, Boot |
|---|---|---|---|---|---|---|---|---|---|
| 1 | VM | Fully Emulated | | VS | VS | VS | VS | VS | VS |
| 2 | VM | Xen PV 3.0 | PV drivers | P | P | P | P | VS | VS |
| 3 | VM | Xen HVM 3.0 | PV drivers | VH | P | P | P | VS | VS |
| 4 | VM | Xen HVM 4.0.1 | PVHVM drivers | VH | P | P | P | P | VS |
| 5 | VM | Xen AWS 2013 | PVHVM + SR-IOV(net) | VH | VH | P | P | P | VS |
| 6 | VM | Xen AWS 2017 | PVHVM + SR-IOV(net, stor.) | VH | VH | VH | P | P | VS |
| 7 | VM | AWS Nitro 2017 | | VH | VH | VH | VH | VH | VS |
| 8 | HW | AWS Bare Metal 2017 | | H | H | H | H | H | H |
| | | Bare Metal | | H | H | H | H | H | H |

VM: Virtual Machine. HW: Hardware.
VS: Virt. in software. VH: Virt. in hardware. P: Paravirt. Not all combinations shown.
SR-IOV(net): ixgbe/ena driver. SR-IOV(storage): nvme driver.

http://www.brendangregg.com/blog/2017-11-29/aws-ec2-virtualization-2017.html

Source: [Gregg 17]

# Lightweight VMs




Source: Systems Performance 2nd Edition, Figure 11.4 [Gregg 20]

Examples:

- AWS "**Firecracker**"
- Intel/ARM/AMD/Microsoft/etc. "**Cloud Hypervisor**" [CloudHypervisor 22]

# My Prediction: Containers

## Perf tools take several years to be fully “container aware”

- Includes non-root BPF work.
- It’s a lot of work, and not enough engineers are working on it. We’ll use workarounds in the meantime (e.g., Kyle Anderson and Sargun Dhillon have made perf tools work in containers at Netflix).
- Was the same with Solaris Zones (long slow process).

# My Prediction: Landscape

Short term:

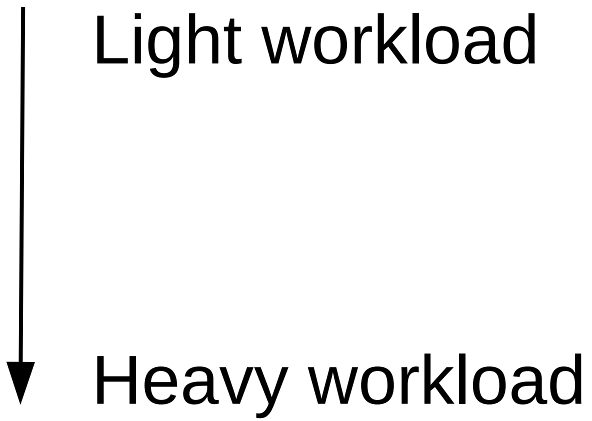
- Containers everywhere

Long term:

- More containers than VMs
- More lightweight VM cores than container cores
  - Hottest workloads switch to dedicated kernels (**no kernel resource sharing, no seccomp overhead, no overlay overhead, full perf tool access, PGO kernels, etc.**)

# My Prediction: Evolution

1. FaaS
2. Container
3. Lightweight VM
4. Metal




Many apps aren't heavy
Metal can also mean single container on metal

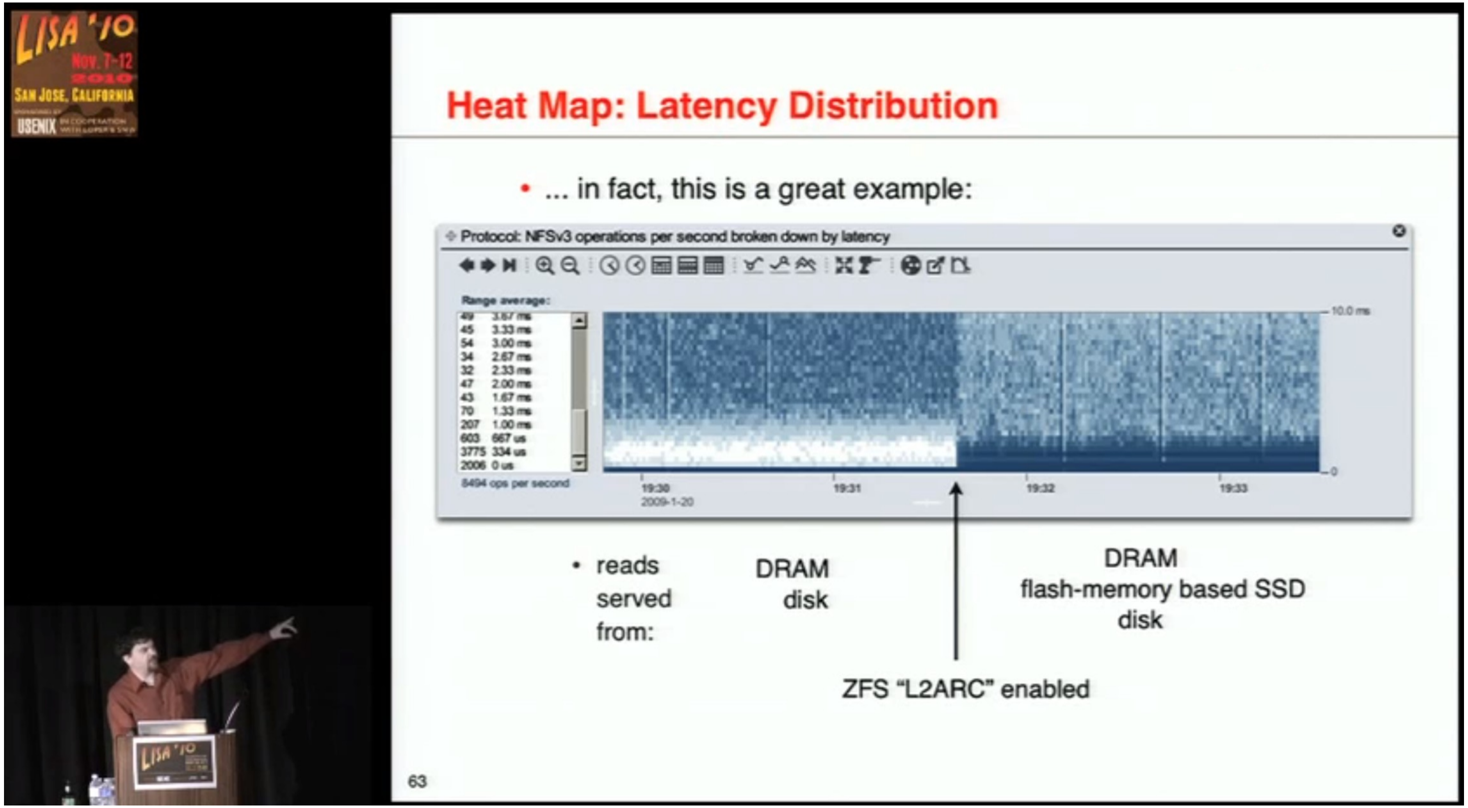
# My Prediction: Cloud Computing

Microservice consolidation becomes a hot topic, to lower communication costs

- **Container schedulers** co-locating chatty services
  - With BPF-based accelerated networking between them (e.g., Cilium)
- Cloud-wide **runtime schedulers** co-locating apps
  - Multiple apps under one JVM roof and process address space

# 7. Observability

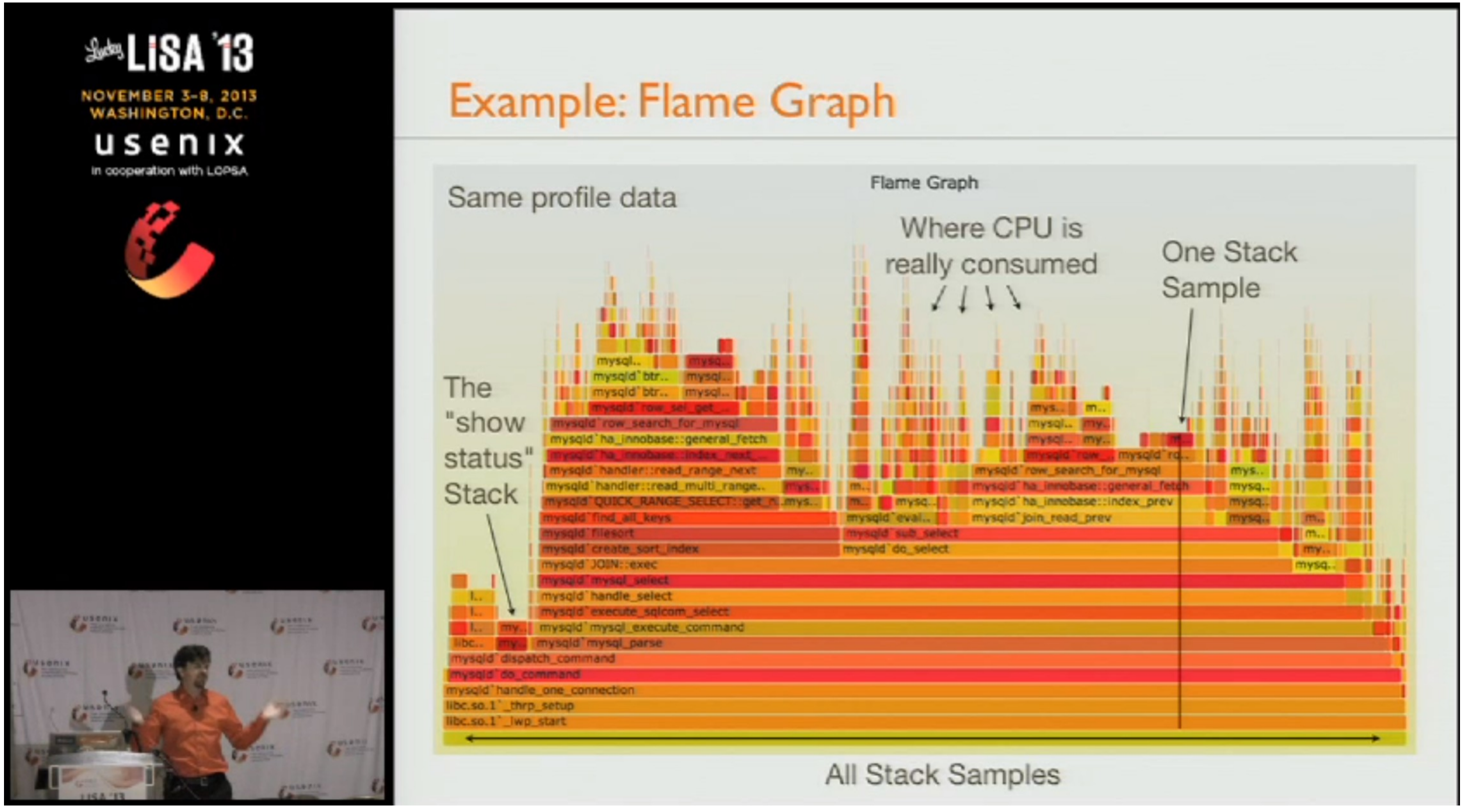
# USENIX 2010: Heat maps




[Gregg 10]

2022: Latency heat maps everywhere

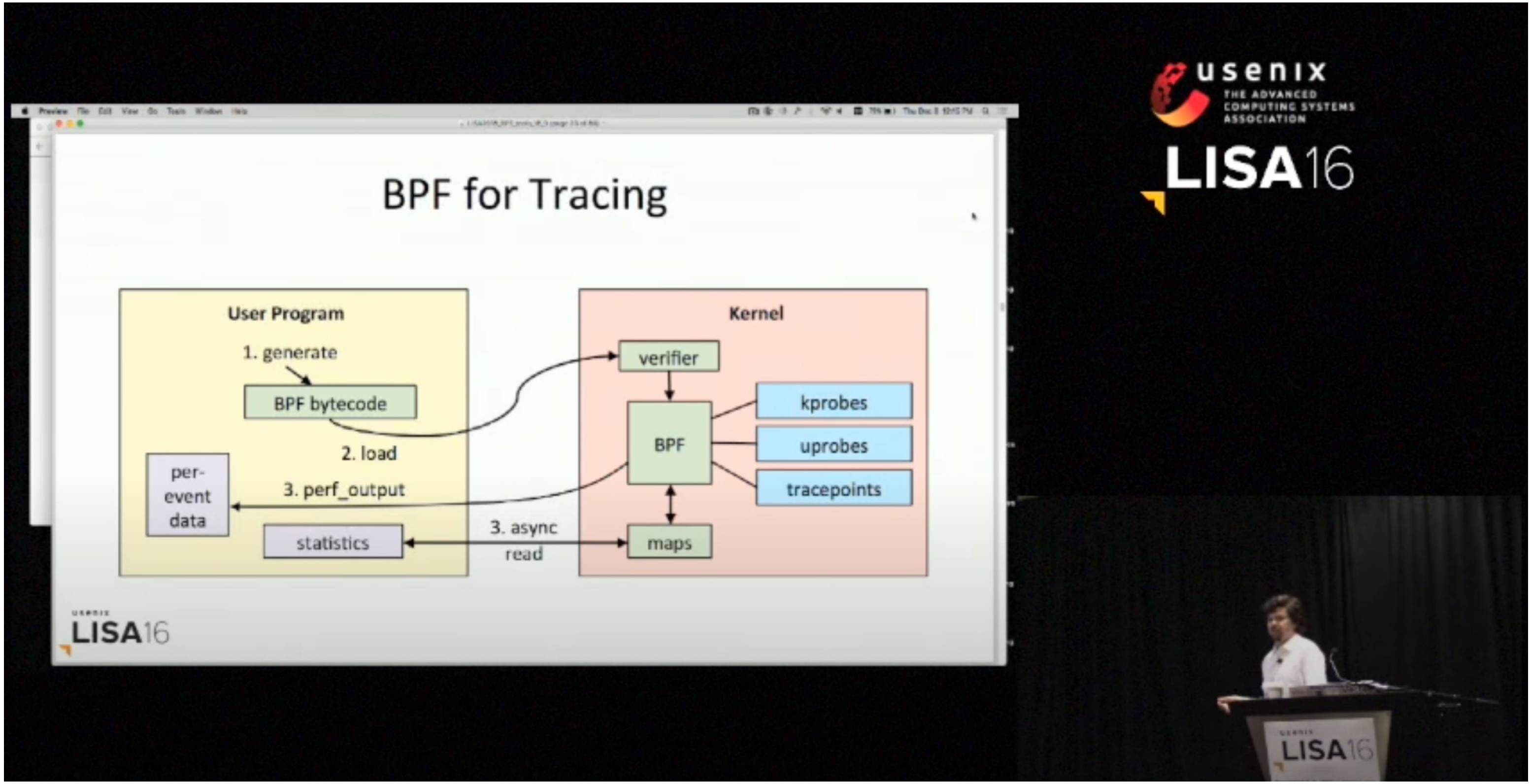
# USENIX 2013: Flame graphs




[Gregg 13]

2022: Flame graphs everywhere

# USENIX 2016: BPF

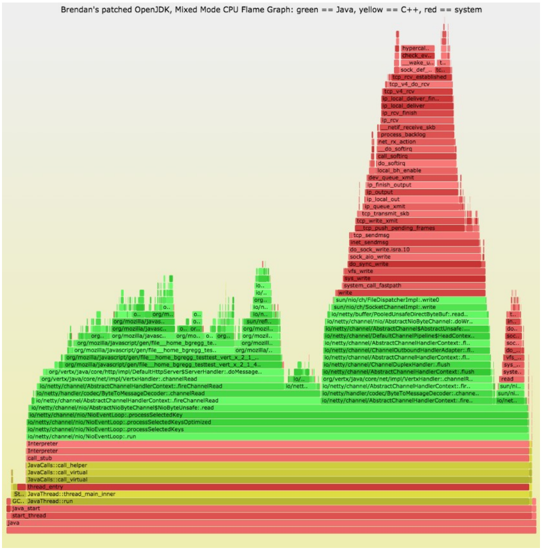


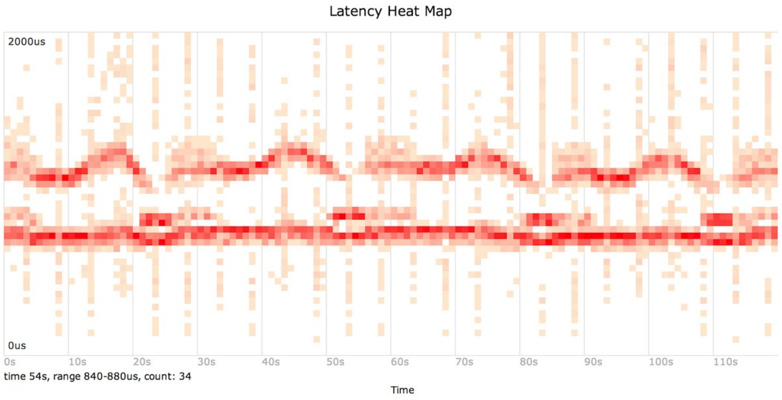

[Gregg 16b]

2022: BPF heading everywhere

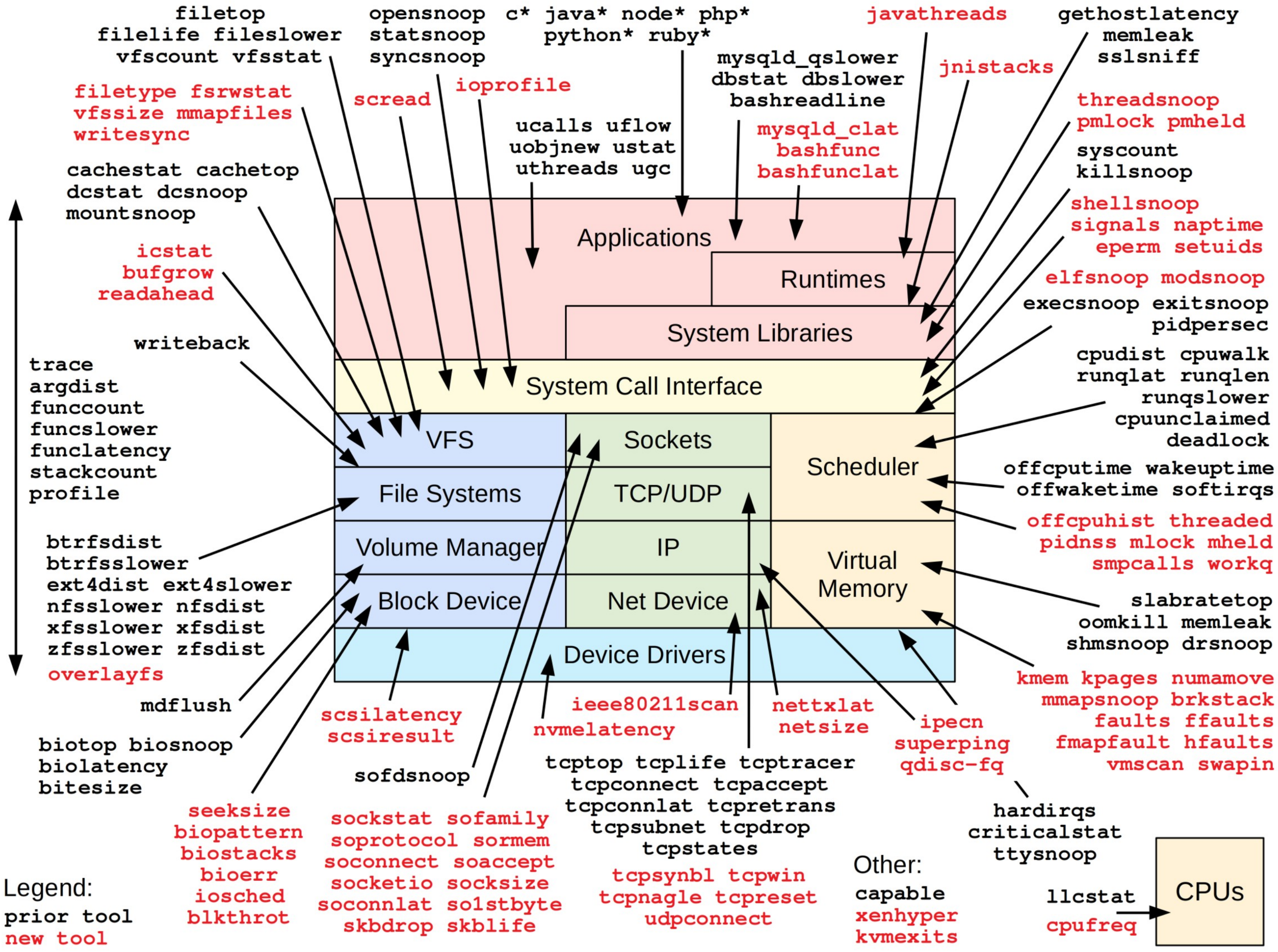
# 2022: Age of Seeing

- Flame graphs everywhere
- Latency heat maps
- eBPF & bpftrace
- PMCs in the cloud



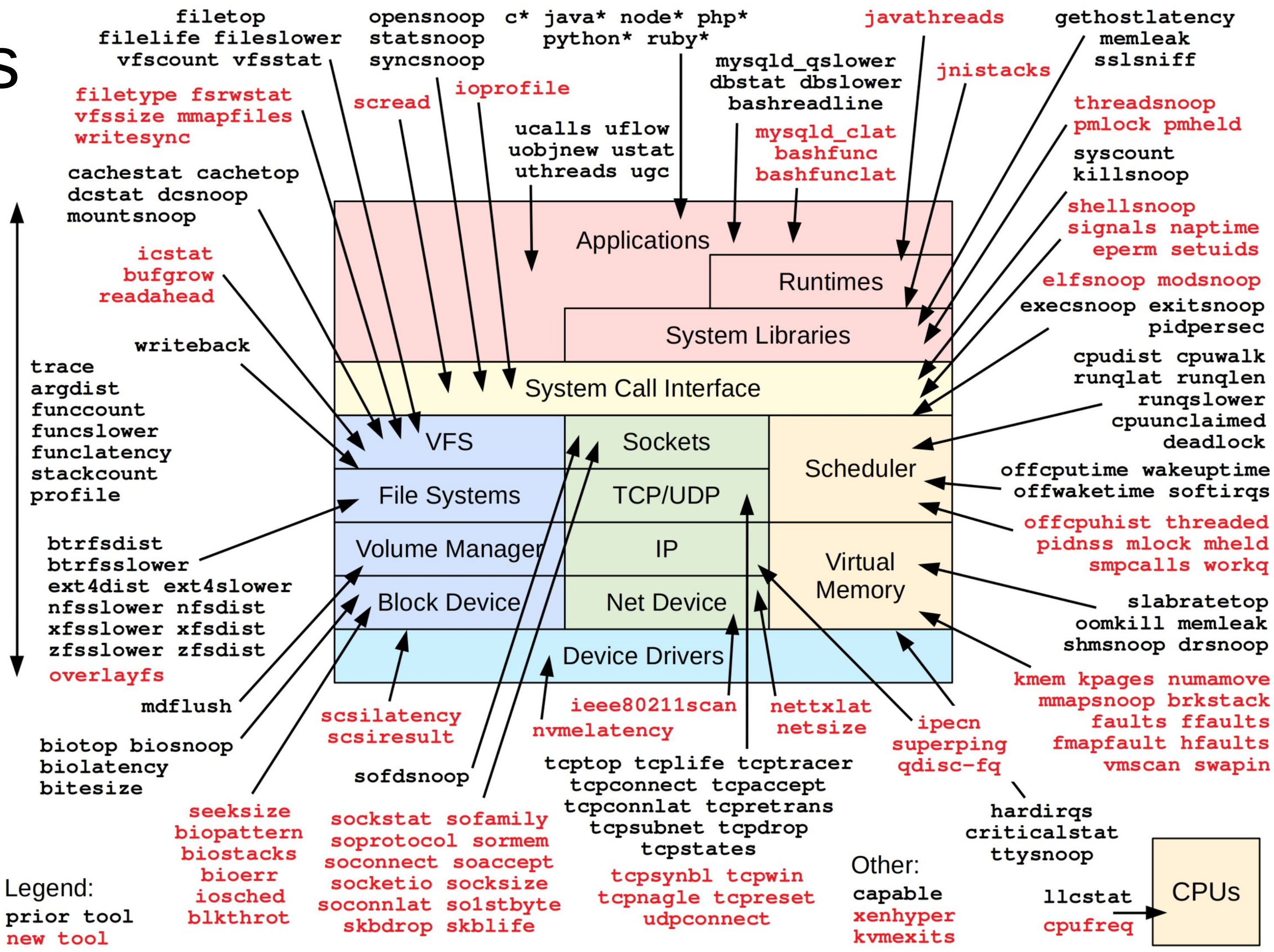



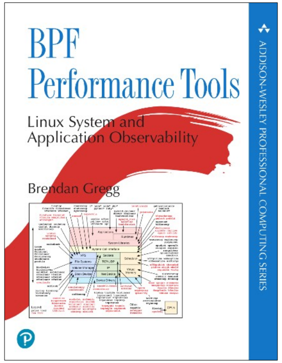

More info: flame graphs [Gregg 13], heat maps [Gregg 10], and eBPF [Gregg 16b]


filetop
filelife fileslower
vfscount vfsstat
opensnoop
statsnoop
syncsnoop
c* java* node* php*
python* ruby*
javathreads
gethostlatency
memleak
sslsniff
filetype fsrwstat
vfssize mmapfiles
writesync
scread
ioprofile
mysqld_qslower
dbstat dbslower
bashreadline
jnistacks
threadsnoop
pmlock pmheld
ucalls uflow
uobjnew ustat
uthreads ugc
mysqld_clat
bashfunc
bashfunclat
syscount
killsnoop
cachestat cachetop
dcstat dcsnoop
mountsnoop
shellsnoop
signals naptime
eperm setuids
icstat
bufgrow
readahead
Applications
Runtimes
System Libraries
elfsnoop modsnoop
execsnoop exitsnoop
pidpersec
writeback
trace
argdist
funccount
funcslower
funclatency
stackcount
profile
System Call Interface
cpudist cpuwalk
runqlat runqlen
runqslower
cpuunclaimed
deadlock
VFS
Sockets
Scheduler
File Systems
TCP/UDP
offcputime wakeuptime
offwaketime softirqs
offcpuhist threaded
pidnss mlock mheld
smpcalls workq
btrfsdist
btrfsslower
ext4dist ext4slower
nfsslower nfsdist
xfsslower xfsdist
zfsslower zfsdist
overlayfs
Volume Manager
IP
Virtual Memory
Block Device
Net Device
slabratetop
oomkill memleak
shmsnoop drsnoop
Device Drivers
kmem kpages numamove
mmapsnoop brkstack
faults ffaults
fmapfault hfaults
vmscan swapin
mdflush
scsilatency
scsiresult
ieee80211scan
nvmelatency
nettxlat
netsize
ipecn
superping
qdisc-fq
biotop biosnoop
biolatency
bitesize
sofdsnoop
tcptop tcplife tcptracer
tcpconnect tcpaccept
tcpconnlat tcpretrans
tcpsubnet tcpdrop
tcpstates
hardirqs
criticalstat
ttysnoop
seeksize
biopattern
biostacks
bioerr
iosched
blkthrot
sockstat sofamily
soprotocol sormem
soconnect soaccept
socketio socksize
soconnlat so1stbyte
skbdrop skblife
tcpsynbl tcpwin
tcpnagle tcpreset
udpconnect
Other:
capable
xenhyper
kvmexits
llcstat
cpufreq
CPUs
Legend:
prior tool
new tool

# BPF Perf Tools

(In red are the new open source tools I developed for the BPF book)



# Example BPF tool

```
# execsnoop.py -T
TIME(s) PCOMM            PID    PPID   RET ARGS
0.506   run              8745   1828     0 ./run
0.507   bash             8745   1828     0 /bin/bash
0.511   svstat           8747   8746     0 /command/svstat /service/httpd
0.511   perl             8748   8746     0 /usr/bin/perl -e $l=<>;$l=~/(\d+) sec/;p...
0.514   ps               8750   8749     0 /bin/ps --ppid 1 -o pid,cmd,args
0.514   grep             8751   8749     0 /bin/grep org.apache.catalina
0.514   sed              8752   8749     0 /bin/sed s/^ *//;
0.515   xargs            8754   8749     0 /usr/bin/xargs
0.515   cut              8753   8749     0 /usr/bin/cut -d  -f 1
0.523   echo             8755   8754     0 /bin/echo
0.524   mkdir            8756   8745     0 /bin/mkdir -v -p /data/tomcat
[...]
1.528   run              8785   1828     0 ./run
1.529   bash             8785   1828     0 /bin/bash
1.533   svstat           8787   8786     0 /command/svstat /service/httpd
1.533   perl             8788   8786     0 /usr/bin/perl -e $l=<>;$l=~/(\d+) sec/;p...
[...]
```

# Example bpftrace one-liner

```
# bpftrace -e 't:block:block_rq_issue { @[args->rwbs] = count(); }'
Attaching 1 probe...
^C

@[R]: 1
@[RM]: 1
@[WFS]: 2
@[FF]: 3
@[WSM]: 9
@[RA]: 10
@[WM]: 12
@[WS]: 29
@[R]: 107
```

# libbpf-tools

```
# ./opensnoop
PID    COMM              FD ERR PATH
27974  opensnoop         28   0 /etc/localtime
1482   redis-server       7   0 /proc/1482/stat
[…]

# ldd opensnoop
    linux-vdso.so.1 (0x00007ffddf3f1000)
    libelf.so.1 => /usr/lib/x86_64-linux-gnu/libelf.so.1 (0x00007f9fb7836000)
    libz.so.1 => /lib/x86_64-linux-gnu/libz.so.1 (0x00007f9fb7619000)
    libc.so.6 => /lib/x86_64-linux-gnu/libc.so.6 (0x00007f9fb7228000)
    /lib64/ld-linux-x86-64.so.2 (0x00007f9fb7c76000)

# ls -lh opensnoop opensnoop.stripped
-rwxr-xr-x 1 root root 645K Feb 28 23:18 opensnoop
-rwxr-xr-x 1 root root 151K Feb 28 23:33 opensnoop.stripped
```

- 151 Kbytes for a stand-alone BPF program!
- (Note: A static bpftrace/BTF + scripts will also have a small average tool size)

# Modern Open Source Observability Stack

## OpenTelemetry

- Standard for monitoring and tracing

## Prometheus

- Monitoring database

## Grafana

- UI with dashboards
- Now supports flame graphs [GrafanaLabs 22]

Grafana


Source: Figure 1.4 [Gregg 20]

# Zero-Instrumentation APM

(Application Performance Monitoring)

## Installation instructions:

1) Install the agent
2) Done! (no code changes required)

## Uses uprobes to instrument HTTP/SSL calls

- Don't even need to restart anything. Great for apps where you can't change the code. But,
- uprobes are slow and unstable: >1.2us minimum, which is 15x higher than kprobes.
- I would not recommend this approach until:
  - Someone does the Linux uprobe speedup work I discussed at LSFMMBPF22 in Palm Springs.
  - USDT is provided instead (which is based on uprobes) to fix stability.
- Can also use eBPF for in-kernel aggregations and programs.

# My Prediction: BPF tool front-ends

## bpftrace

- For one-liners and to hack up new tools
- When you want to spend an afternoon developing some custom BPF tracing

## libbpf-tools

- For packaged BPF binary tools and BPF products
- When you want to spend weeks developing BPF

# My Prediction: Too many BPF tools

(I'm partly to blame)

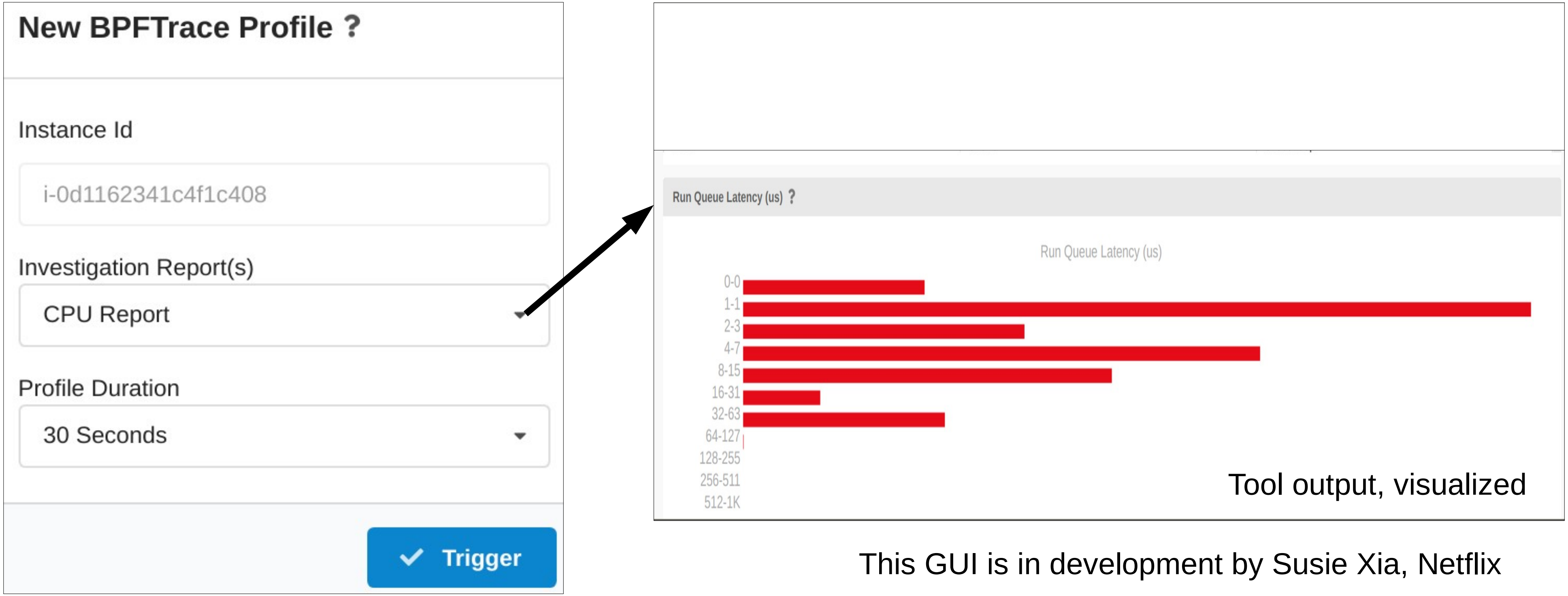
2014: I have **no tools** for this problem

20**24**: I have **too many tools** for this problem

Tool creators: Focus on solving something no other tool can. Necessity is the mother of good BPF tools.

# My Prediction: BPF perf tool future

## GUIs, not CLI tools




This GUI is in development by Susie Xia, Netflix

The end user may not even know it's using BPF

# My Prediction: Zero-instrumentation APM

Multiple startups will be selling this

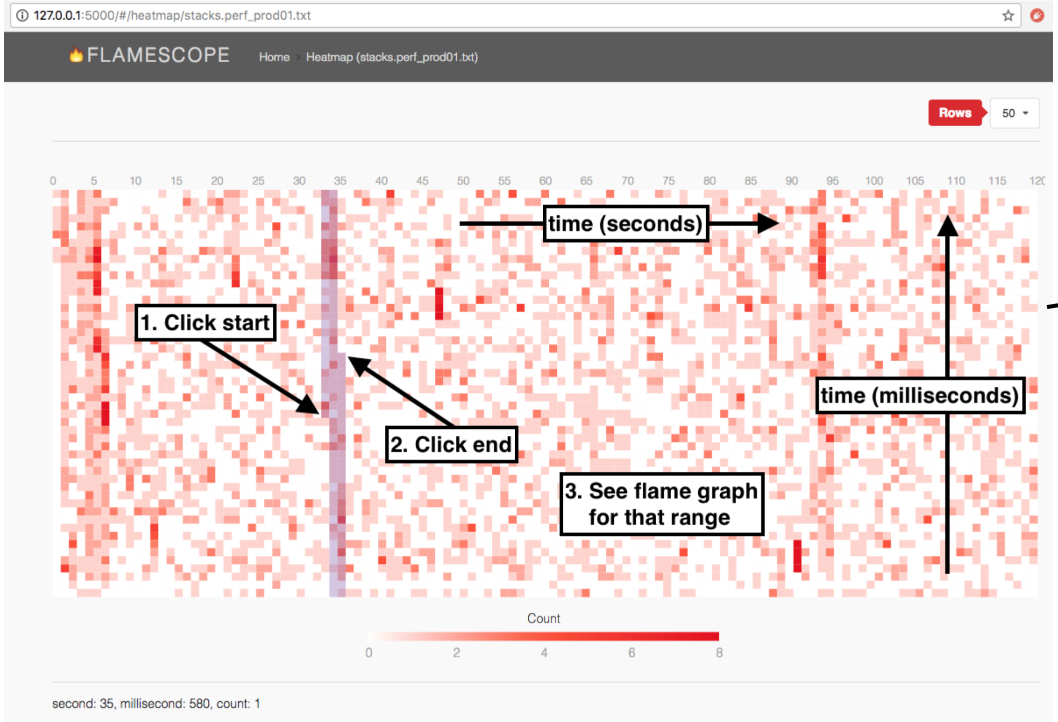
Someone blogs: "OpenTelemetry more stable *and faster*"

- This gives uprobes/eBPF a bad name, unfairly, as none of us in uprobe/eBPF land recommend this use case until the speed/stability issues are fixed

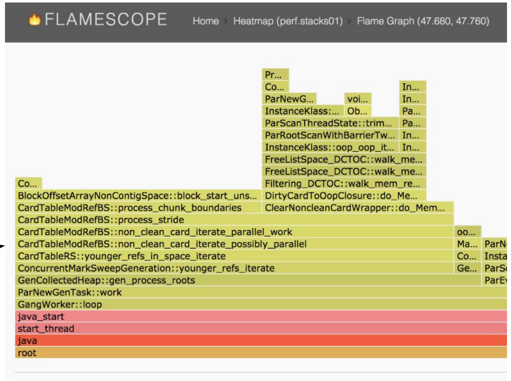
Fast uprobes available in Linux in 2024?

# My Prediction: Flame scope adoption

## Analyze variance, perturbations:




Subsecond-offset heat map




Flame graph

[Spier 20]

# Recap so far

1. Processors
2. Memory
3. Disks
4. Networking
5. Kernels
6. Hypervisors
7. Observability

# Performance engineering is getting more **complex**

1. Processors: **CPUs, GPUs, FPGAs, TPUs**
2. Memory: **DRAM, RLDRAM, HBM, 3D XPoint**
3. Disks: **PMR, SMR, MAT, EAMR, HAMR, SLC, MLC, ...**
4. Networking: **QUIC, MP-TCP, XDP, qdiscs, pacing, BQL, ...**
5. Kernels: **BPF, io_uring, PGO, Linux complexity**
6. Hypervisors: **VMs, Containers, LightweightVMs**
7. Observability: **BPF, PMCs, heat maps, flame graphs, OpenTelemetry, Prometheus, Grafana**

# Performance engineering is getting more **fun**!

1. Processors: **CPUs, GPUs, FPGAs, TPUs**
2. Memory: **DRAM, RLDRAM, HBM, 3D XPoint**
3. Disks: **PMR, SMR, MAT, EAMR, HAMR, SLC, MLC, ...**
4. Networking: **QUIC, MP-TCP, XDP, qdiscs, pacing, BQL, ...**
5. Kernels: **BPF, io_uring, PGO, Linux complexity**
6. Hypervisors: **VMs, Containers, LightweightVMs**
7. Observability: **BPF, PMCs, heat maps, flame graphs, OpenTelemetry, Prometheus, Grafana**

# 8. AI

# AI Auto-Tuning

## One approach to deal with the complexity:

All The Things!

1. Processors: **CPUs, GPUs, FPGAs, TPUs**
2. Memory: **DRAM, RLDRAM, HBM, 3D XPoint**
3. Disks: **PMR, SMR, MAT, EAMR, HAMR, SLC, MLC, ...**
4. Networking: **QUIC, MP-TCP, XDP, qdiscs, pacing, BQL, ...**
5. Kernels: **BPF, io_uring, PGO, Linux complexity**
6. Hypervisors: **VMs, Containers, LightweightVMs**
7. Observability: **BPF, PMCs, heat maps, flame graphs, OpenTelemetry, Prometheus, Grafana**

→ AI → Server tuning

E.g., Intel Granulate,
Akamas,
etc.

## Implications for SRE? (E.g., change control)

# My Prediction: AI useful but limited

- Great at adapting the past, applying known tuning
  - Paint an astronaut in the style of Van Gogh
  - Apply system tuning in the style of Brendan Gregg (tunables I've shared in the past)
  - Should be a useful and more widely-adopted product, especially for small/medium sites that have little time for tuning.
- Turning point: "The Million Dollar Tunable"
  - Someone uses AI to find an overlooked tunable that they coud have enabled years ago.
  - The industry is getting more complex, and more chances things are overlooked. The time is right for using AI to help.
- Poor at solving "never seen before" mental-leap issues
  - I spend most of my time as a performance engineer solving these
  - Beyond past experience or extrapolation

# References (1)


[Gregg 08] Brendan Gregg, “ZFS L2ARC,” http://www.brendangregg.com/blog/2008-07-22/zfs-l2arc.html, Jul 2008

[Gregg 10] Brendan Gregg, “Visualizations for Performance Analysis (and More),” https://www.usenix.org/conference/lisa10/visualizations-performance-analysis-and-more, 2010

[Greenberg 11] Marc Greenberg, “DDR4: Double the speed, double the latency? Make sure your system can handle next-generation DRAM,” https://www.chipestimate.com/DDR4-Double-the-speed-double-the-latencyMake-sure-your-system-can-handle-next-generation-DRAM/Cadence/Technical-Article/2011/11/22, Nov 2011

[Hruska 12] Joel Hruska, “The future of CPU scaling: Exploring options on the cutting edge,” https://www.extremetech.com/computing/184946-14nm-7nm-5nm-how-low-can-cmos-go-it-depends-if-you-ask-the-engineers-or-the-economists, Feb 2012

[Gregg 13] Brendan Gregg, “Blazing Performance with Flame Graphs,” https://www.usenix.org/conference/lisa13/technical-sessions/plenary/gregg, 2013

[Shimpi 13] Anand Lal Shimpi, “Seagate to Ship 5TB HDD in 2014 using Shingled Magnetic Recording,” https://www.anandtech.com/show/7290/seagate-to-ship-5tb-hdd-in-2014-using-shingled-magnetic-recording, Sep 2013

[Borkmann 14] Daniel Borkmann, “net: tcp: add DCTCP congestion control algorithm,” https://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git/commit/?id=e3118e8359bb7c59555aca60c725106e6d78c5ce, 2014

[Macri 15] Joe Macri, “Introducing HBM,” https://www.amd.com/en/technologies/hbm, Jul 2015

[Cardwell 16] Neal Cardwell, et al., “BBR: Congestion-Based Congestion Control,” https://queue.acm.org/detail.cfm?id=3022184, 2016

[Gregg 16] Brendan Gregg, “Unikernel Profiling: Flame Graphs from dom0,” http://www.brendangregg.com/blog/2016-01-27/unikernel-profiling-from-dom0.html, Jan 2016

# References (2)


[Gregg 16b] Brendan Gregg, “Linux BPF Superpowers,” https://www.brendangregg.com/blog/2016-03-05/linux-bpf-superpowers.html, 2016

[Alcorn 17] Paul Alcorn, “Seagate To Double HDD Speed With Multi-Actuator Technology,” https://www.tomshardware.com/news/hdd-multi-actuator-heads-seagate,36132.html, 2017

[Alcorn 17b] Paul Alcorn, “Hot Chips 2017: Intel Deep Dives Into EMIB,” https://www.tomshardware.com/news/intel-emib-interconnect-fpga-chiplet,35316.html#xenforo-comments-3112212, 2017

[Corbet 17] Jonathan Corbet, “Two new block I/O schedulers for 4.12,” https://lwn.net/Articles/720675, Apr 2017

[Gregg 17] Brendan Gregg, “AWS EC2 Virtualization 2017: Introducing Nitro,” http://www.brendangregg.com/blog/2017-11-29/aws-ec2-virtualization-2017.html, Nov 2017

[Russinovich 17] Mark Russinovich, “Inside the Microsoft FPGA-based configurable cloud,” https://www.microsoft.com/en-us/research/video/inside-microsoft-fpga-based-configurable-cloud, 2017

[Gregg 18] Brendan Gregg, “Linux Performance 2018,” http://www.brendangregg.com/Slides/Percona2018_Linux_Performance.pdf, 2018

[Hady 18] Frank Hady, “Achieve Consistent Low Latency for Your Storage-Intensive Workloads,” https://www.intel.com/content/www/us/en/architecture-and-technology/optane-technology/low-latency-for-storage-intensive-workloads-article-brief.html, 2018

[Joshi 18] Amit Joshi, et al., “Titus, the Netflix container management platform, is now open source,” https://netflixtechblog.com/titus-the-netflix-container-management-platform-is-now-open-source-f868c9fb5436, Apr 2018

[Cutress 19] Dr. Ian Cutress, “Xilinx Announces World Largest FPGA: Virtex Ultrascale+ VU19P with 9m Cells,” https://www.anandtech.com/show/14798/xilinx-announces-world-largest-fpga-virtex-ultrascale-vu19p-with-9m-cells, Aug 2019

# References (3)


[Gallatin 19] Drew Gallatin, “Kernel TLS and hardware TLS offload in FreeBSD 13,” https://people.freebsd.org/~gallatin/talks/euro2019-ktls.pdf, 2019

[Bearman 20] Ian Bearman, “Exploring Profile Guided Optimization of the Linux Kernel,” https://linuxplumbersconf.org/event/7/contributions/771, 2020

[Burnes 20] Andrew Burnes, “GeForce RTX 30 Series Graphics Cards: The Ultimate Play,” https://www.nvidia.com/en-us/geforce/news/introducing-rtx-30-series-graphics-cards, Sep 2020

[Charlene 20] Charlene, “800G Is Coming: Set Pace to More Higher Speed Applications,” https://community.fs.com/blog/800-gigabit-ethernet-and-optics.html, May 2020

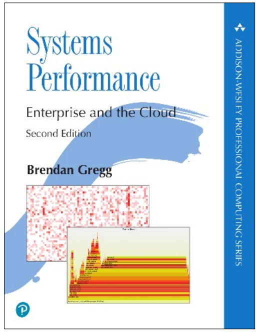
[Cutress 20] Dr. Ian Cutress, “Insights into DDR5 Sub-timings and Latencies,” https://www.anandtech.com/show/16143/insights-into-ddr5-subtimings-and-latencies, Oct 2020

[Ford 20] A. Ford, et al., “TCP Extensions for Multipath Operation with Multiple Addresses,” https://datatracker.ietf.org/doc/html/rfc8684, Mar 2020

[Gregg 20] Brendan Gregg, “Systems Performance: Enterprise and the Cloud, Second Edition,” *Addison-Wesley*, 2020

[Hruska 20] Joel Hruska, “Intel Demos PCIe 5.0 on Upcoming Sapphire Rapids CPUs,” https://www.extremetech.com/computing/316257-intel-demos-pcie-5-0-on-upcoming-sapphire-rapids-cpus, Oct 2020

[Liu 20] Linda Liu, “Samsung QVO vs EVO vs PRO: What’s the Difference? [Clone Disk],” https://www.partitionwizard.com/clone-disk/samsung-qvo-vs-evo.html, 2020

# References (4)


[Moore 20] Samuel K. Moore, “A Better Way to Measure Progress in Semiconductors,” https://spectrum.ieee.org/semiconductors/devices/a-better-way-to-measure-progress-in-semiconductors, Jul 2020

[Peterson 20] Zachariah Peterson, “DDR5 vs. DDR6: Here's What to Expect in RAM Modules,” https://resources.altium.com/p/ddr5-vs-ddr6-heres-what-expect-ram-modules, Nov 2020

[Salter 20] Jim Salter, “Western Digital releases new 18TB, 20TB EAMR drives,” https://arstechnica.com/gadgets/2020/07/western-digital-releases-new-18tb-20tb-eamr-drives, Jul 2020

[Spier 20] Martin Spier, Brendan Gregg, et al., “FlameScope,” https://github.com/Netflix/flamescope, 2020

[Tolvanen 20] Sami Tolvanen, Bill Wendling, and Nick Desaulniers, “LTO, PGO, and AutoFDO in the Kernel,” Linux Plumber’s Conference, https://linuxplumbersconf.org/event/7/contributions/798, 2020

[Vega 20] Juan Camilo Vega, Marco Antonio Merlini, Paul Chow, “FFShark: A 100G FPGA Implementation of BPF Filtering for Wireshark,” *IEEE 28th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2020

[Warren 20] Tom Warren, “Microsoft reportedly designing its own ARM-based chips for servers and Surface PCs,” https://www.theverge.com/2020/12/18/22189450/microsoft-arm-processors-chips-servers-surface-report, Dec 2020

[Alcorn 21] Paul Alcorn, “Intel Shares Alder Lake Pricing, Specs and Gaming Performance: $589 for 16 Cores,” https://www.tomshardware.com/features/intel-shares-alder-lake-pricing-specs-and-gaming-performance, Oct 2021

[Cutress 21] Ian Cutress, “AMD Demonstrates Stacked 3D V-Cache Technology: 192 MB at 2 TB/sec,” https://www.anandtech.com/show/16725/amd-demonstrates-stacked-vcache-technology-2-tbsec-for-15-gaming, May 2021

[Google 21] Google, “Cloud TPU,” https://cloud.google.com/tpu, 2021

# References (5)


[Haken 21] Michael Haken, et al., “Delta Lake 1S Server Design Specification 1v05, https://www.opencompute.org/documents/delta-lake-1s-server-design-specification-1v05-pdf, 2021

[Intel 21] Intel corporation, “Intel® OptaneTM Technology,” https://www.intel.com/content/www/us/en/products/docs/storage/optane-technology-brief.html, 2021

[Kostovic 21] Aleksandar Kostovic, “Esperanto Delivers Kilocore Processor in its Supercomputer-on-a-Chip Design,” https://www.tomshardware.com/news/esperanto-kilocore-processor, Aug 2021

[Kummrow 21] Patricia Kummrow, “The IPU: A New, Strategic Resource for Cloud Service Providers,” https://itpeernetwork.intel.com/ipu-cloud/#gs.g5pkub, Aug 2021

[Quach 21a] Katyanna Quach, “Global chip shortage probably won't let up until 2023, warns TSMC: CEO 'still expects capacity to tighten more',” https://www.theregister.com/2021/04/16/tsmc_chip_forecast, Apr 2021

[Quach 21b] Katyanna Quach, “IBM says it's built the world's first 2nm semiconductor chips,” https://www.theregister.com/2021/05/06/ibm_2nm_semiconductor_chips, May 2021

[Ridley 21] Jacob Ridley, “IBM agrees with Intel and TSMC: this chip shortage isn't going to end anytime soon,” https://www.pcgamer.com/ibm-agrees-with-intel-and-tsmc-this-chip-shortage-isnt-going-to-end-anytime-soon, May 2021

[Shilov 21] Anton Shilov, “Samsung Develops 512GB DDR5 Module with HKMG DDR5 Chips,” https://www.tomshardware.com/news/samsung-512gb-ddr5-memory-module, Mar 2021

[Shilov 21b] Anton Shilov, “Seagate Ships 20TB HAMR HDDs Commercially, Increases Shipments of Mach.2 Drives,” https://www.tomshardware.com/news/seagate-ships-hamr-hdds-increases-dual-actuator-shipments, 2021

[Shilov 21c] Anton Shilov, “SK Hynix Envisions 600-Layer 3D NAND & EUV-Based DRAM,” https://www.tomshardware.com/news/sk-hynix-600-layer-3d-nand-euv-dram, Mar 2021

# References (6)


[SuperMicro 21] SuperMicro, “B12SPE-CPU-25G (For SuperServer Only),” https://www.supermicro.com/en/products/motherboard/B12SPE-CPU-25G, 2021

[Thaler 21] Dave Thaler, Poorna Gaddehosur, “Making eBPF work on Windows,” https://cloudblogs.microsoft.com/opensource/2021/05/10/making-ebpf-work-on-windows, May 2021

[TornadoVM 21] TornadoVM, “TornadoVM Run your software faster and simpler!” https://www.tornadovm.org, 2021

[Trader 21] Tiffany Trader, “Cerebras Second-Gen 7nm Wafer Scale Engine Doubles AI Performance Over First-Gen Chip ,” https://www.enterpriseai.news/2021/04/21/latest-cerebras-second-gen-7nm-wafer-scale-engine-doubles-ai-performance-over-first-gen-chip, Apr 2021

[Ghigoff 21] Yoann Ghigoff, et al., "BMC: Accelerating Memcached using Safe In-kernel Caching and Pre-stack Processing," Proceedings of the 18th USENIX Symposium on Networked Systems Design and Implementation, https://www.usenix.org/system/files/nsdi21-ghigoff.pdf, Apr 2021

[Tyson 21] Mark Tyson, "Intel Sapphire Rapids utillises tiled, modular SoC architecture," https://hexus.net/tech/news/cpu/148266-intel-sapphire-rapids-utillises-tiled-modular-soc-architecture/, Aug 2021

[Vahdat 21] Amin Vahdat, “The past, present and future of custom compute at Google,” https://cloud.google.com/blog/topics/systems/the-past-present-and-future-of-custom-compute-at-google, Mar 2021

[Wikipedia 21] “Semiconductor device fabrication,” https://en.wikipedia.org/wiki/Semiconductor_device_fabrication, 2021

[Wikipedia 21b] “Silicon,” https://en.wikipedia.org/wiki/Silicon, 2021

[ZonedStorage 21] Zoned Storage, “Zoned Namespaces (ZNS) SSDs,” https://zonedstorage.io/introduction/zns, 2021

[Cutress 21b] Dr. Ian Cutress, Andrei Frumusanu, "The Intel 12th Gen Core i9-12900K Review: Hybrid Performance Brings Hybrid Complexity," https://www.anandtech.com/show/17047/the-intel-12th-gen-core-i912900k-review-hybrid-performance-brings-hybrid-complexity, Nov 2021

# References (7)


[Nash 22] Paul Nash, "Now in preview: Azure Virtual Machines with Ampere Altra Arm-based processors," https://azure.microsoft.com/en-us/blog/now-in-preview-azure-virtual-machines-with-ampere-altra-armbased-processors/, Apr 2022

[Bonshor 22] Gavin Bonshor, "AMD Releases Milan-X CPUs With 3D V-Cache," https://www.anandtech.com/show/17323/amd-releases-milan-x-cpus-with-3d-vcache-epyc-7003, Mar 2022

[Mann 22] Tobias Mann, "Why Intel killed its Optane memory business," https://www.theregister.com/2022/07/29/intel_optane_memory_dead/, Jul 2022

[Torvalds 22] Linus Torvalds, “Linux 6.0-rc1,” https://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git/commit/?id=568035b01cfb107af8d2e4bd2fb9aea22cf5b868, Aug 2022

[Whalen 22] Jeanne Whalen, "Biden’s visit shows high stakes of $20 billion Ohio chip factory," https://www.washingtonpost.com/us-policy/2022/09/09/biden-intel-ohio-chip-factory/, Sep 2022

[Robinson 22] Dan Robinson, "Intel has a secret club in the cloud for devs to try out new chips – and you ain't in it," https://www.theregister.com/2022/09/28/intel_developer_cloud/, Sep 2022

[CloudHypervisor 22] Cloud Hypervisor Project (Linux Foundation), "Run Cloud Virtual Machines Securely and Efficiently," https://www.cloudhypervisor.org, accessed 2022

[Cerebras 22] Cerebras, "Cerebras Wafer-Scale Cluster," https://www.cerebras.net/product-cluster, accessed 2022

[GrafanaLabs 22] Grafana Labs, "https://grafana.com/docs/grafana/latest/panels-visualizations/visualizations/flame-graph/," https://grafana.com/docs/grafana/latest/panels-visualizations/visualizations/flame-graph/, accessed 2022

[Pirzada 22] Usman Pirzada, "Intel Announces The Worlds First x86 CPU With HBM Memory: Xeon Max ‘Sapphire Rapids’ Data Center CPU," https://wccftech.com/intel-announces-the-worlds-first-x86-cpu-with-hbm-memory-xeon-max-sapphire-rapids-data-center-cpu/, Nov 2022

# References (8)


[Smith 22] Lyle Smith, "Samsung PM1743 PCIe Gen5 SSD First Take Review," https://www.storagereview.com/review/samsung-pm1743-pcie-gen5-ssd-first-take-review, Jan 2022

[Barr 22] Jeff Barr, "New Amazon EC2 Instance Types In the Works – C7gn, R7iz, and Hpc7g," https://aws.amazon.com/blogs/aws/new-amazon-ec2-instance-types-in-the-works-c7gn-r7iz-and-hpc7g, Nov 2022

[Gooding 22] Matthew Gooding, "TSMC's US fab will make 4nm chips for Apple, AMD and Nvidia," https://techmonitor.ai/technology/silicon/tsmcs-arizona-apple-amd-nvidia, Dec 2022

[Liu 22] Zhiye Liu, "Smuggler Hid Over 200 Alder Lake CPUs in Fake Silicone Belly," https://www.tomshardware.com/news/smuggler-hid-over-200-alder-lake-cpus-in-fake-silicone-belly, Dec 2022

[Seagate 22] Seagato, "Exos X Series," https://www.seagate.com/au/en/products/enterprise-drives/exos-x/, accessed 2022

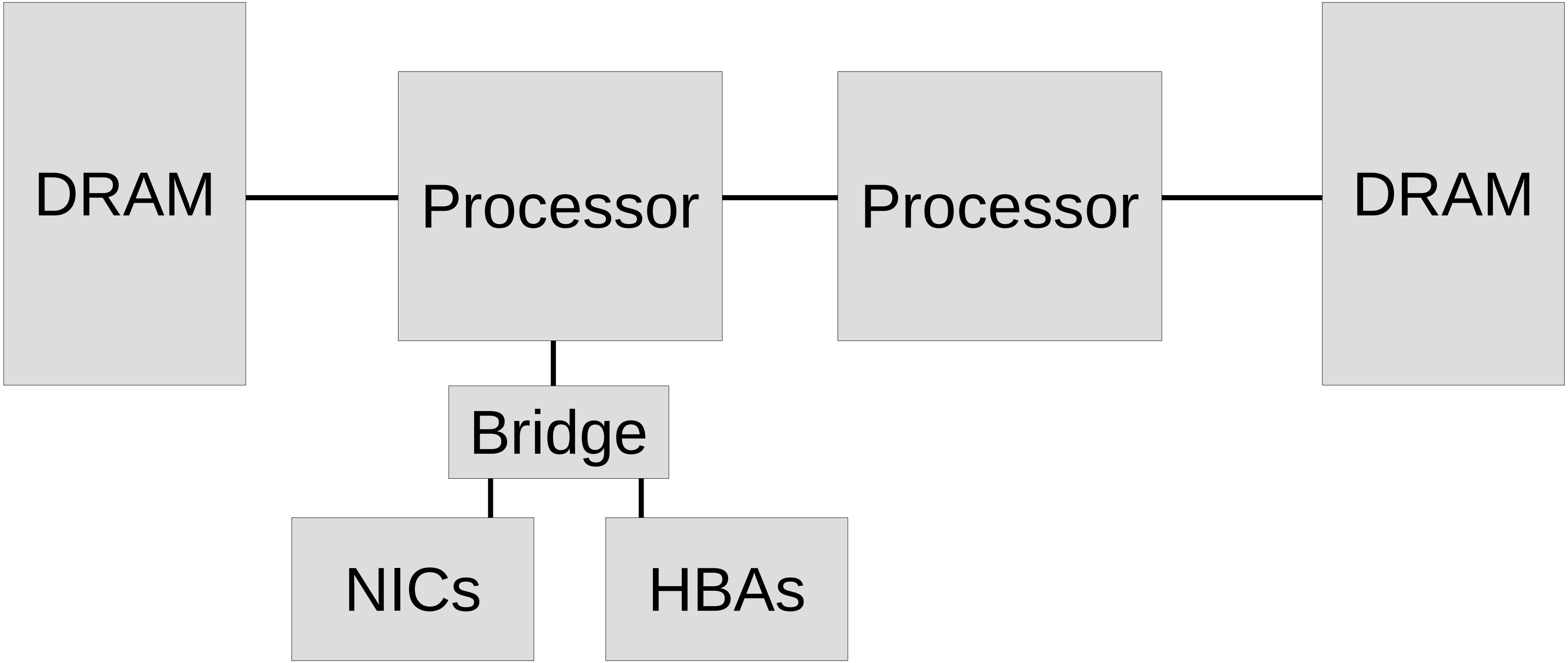
[Mann 22b] Tobais Mann, "Nvidia not cutting it? Google and Amazon's latest AI chips have arrived," https://www.theregister.com/2022/10/11/google_amazon_ai_chips_nvidia/, Oct 2022

[Intel 22] Intel, "Intel® Developer Cloud," https://www.intel.com/content/www/us/en/developer/tools/devcloud/overview.html, accessed Dec 2022

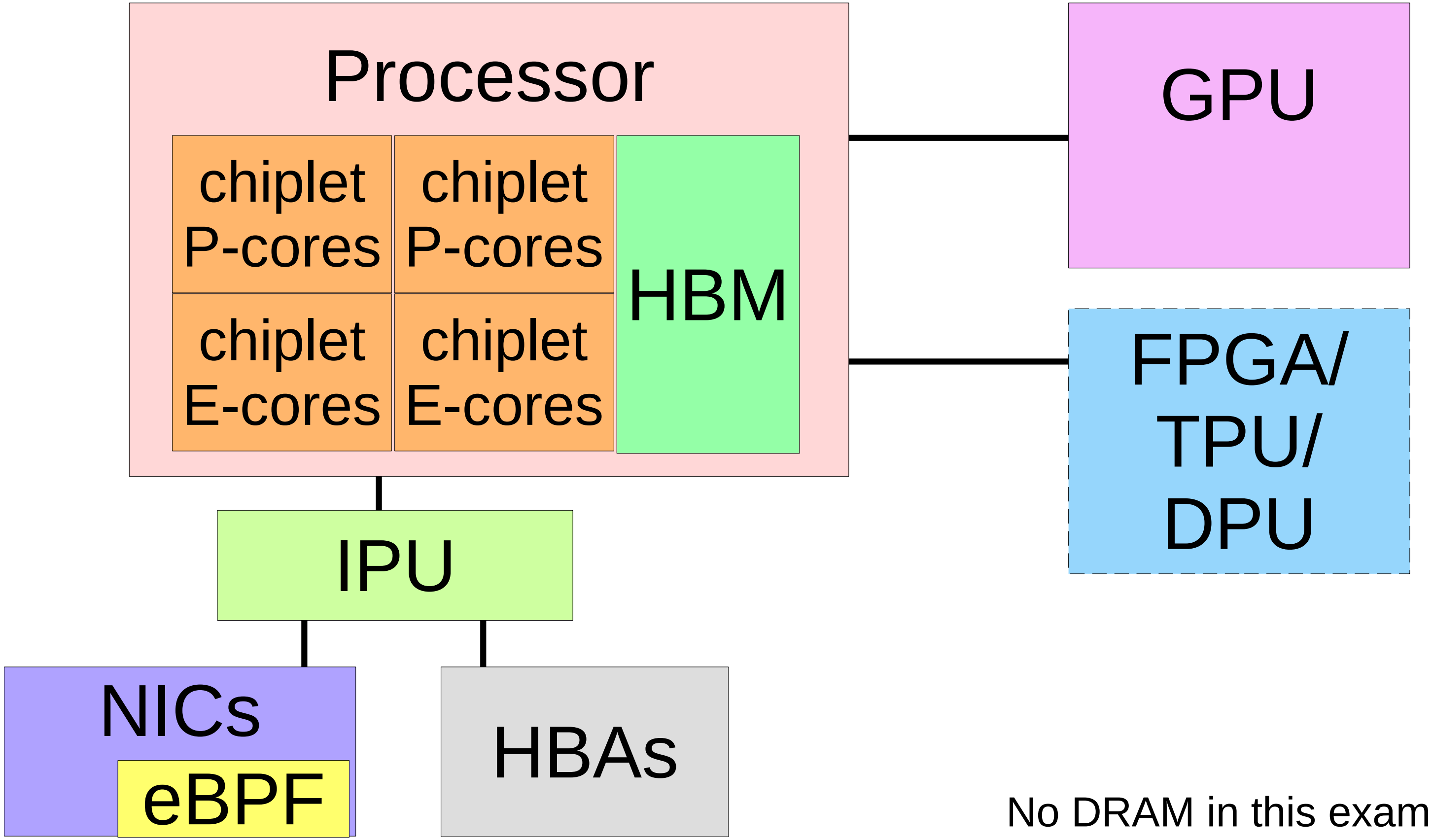
# Take Aways

- Awareness of current and future perf technologies
- Design faster systems to meet SLOs and performance needs
- Begin planning new technology support and maintenance

# Old System



# Example Future System




No DRAM in this example!

# Thanks

Thanks for attending Sydney's first USENIX event!

Slides: http://www.brendangregg.com

Thanks to colleagues Jason Koch, Sargun Dhillon, Drew Gallatin, and Cuy Cirino for their performance engineering expertise.

Thanks to USENIX organizers!

SRECON ASIA PACIFIC

**SYDNEY, AUSTRALIA**
7–9 December, 2022