Why are we here?

• It’s the people!
  – Past attendees include 3 Turing award winners
• My first HPTS was 27 years ago
  – Thanks to Pat Selinger
• Between session discussion vital
• No conference has influenced me as deeply
  – Some conversations have led to as much as a decade of work
Where Have I Been?

• If HPTS so important, where have I been?
• 2012 to 2022 around the world in a small boat
  – Worked full time at AWS
  – Only in North America 3 to 4 times/year
  – Incredible experience
  – Memorable satellite bill :-(
  – Great to be back!
HPTS Comes Full Circle

• Early days of HPTS
  – Purpose-built, vertically-integrated systems
  – HPTS attendees had full control of S/W & H/W
• Horizontally-scaled clusters of commodity servers
  – Scale continues to grow rapidly but little H/W specialization
  – Datacenter becomes the computer
• HPTS today
  – Applications span datacenters for scale & reliability
  – Extraordinary scale drives return to H/W specialization
  – HPTS attendees back to S/W & H/W control
• Full circle examples from AWS
HPTS 1995: 3,692 tps

- My first HPTS in 1995
- 100 million long distance calls/day
  - 1,157 billing recs/sec
- Oracle 6.1: 3,692 tpsA
HPTS 1999

• My first HPTS talk
  – Fault Avoidance vs. Fault Tolerance: Testing Doesn’t Scale

• Core thesis:
  – Transaction systems increasingly complex, distributed, & w/o maintenance windows
  – Fault avoidance ineffective
  – Must be fault tolerant
HPTS 2001: Obidos

• Charlie Bell & Rick Dalzell
• Amazon page rendering engine
  – https://www.amazon.com/exec/obidos/ASIN/0596515162
• One large hairball:
  – 4GB image on 32bit system
  – Frequently broke the GNU linker
  – Leaked memory so quickly that application restarts were required every 100 to 200 requests
• Primary reason Amazon moved to SOA so early
• Lives on as the name of a Seattle office building
• I viewed Obdios as a breakthrough:
  – Highly reliable system composed of highly unreliable parts
• An extreme example of the design direction
  – I loved it and thought it was the future
  – Bruce Lindsey thought it was morally bankrupt engineering
  – We had a great discussion
  – I think we both probably had a point :-)

2022.10.10
http://perspectives.mvdirona.com
HPTS 2005: One Size Fits All

- Thesis: application-specific DBs 10x faster
  - Simple ideas can deliver the most profound impact
  - Unleashed 2 decades of innovation
- At time really only 3 relevant commercial DBs
  - Admin so complex, most customers had only 1 DB
Application-Specific DB

• Cloud computing removes administrative complexity
  – Administration part of the service
  – Much easier to use workload-optimized DBs

• More than 13 unique DB services at AWS
  – Relational: Aurora, MySQL, PostgreSQL, MariaDB
  – Commercial: SQL Server, Oracle
  – DW/Analytics: Redshift, Athena
  – NoSQL: DyamoDB, DocumentDB
  – Graph DB: Neptune
  – In-Memory: ElastiCache, MemoryDB
2009: Datacenter as a Computer

• The industry moves past clusters
  – An entire building of computers under single administrative control
  – Web search is big

• Cloud computing drives scale acceleration
  – Far larger than web search
  – Largest non-gov server count
  – Applications run across datacenters for reliability, scaling, & latency
Cloud Scale Feeds Innovation

• Scale drives R&D, which drives innovation, which drives further scale
  – Gives us back full control of H/W & S/W stack

• Examples from AWS
  – Custom server designs
  – Custom network designs
  – Custom semiconductors
    • Nitro service, storage, security & network offload
    • Graviton server CPU line
    • Inferentia ML inference processor line
    • Trainium ML training processor line
Custom Servers

• AWS has designed & developed custom servers for more than a decade
  – Reduced cost
  – Multi-source contract manufactures
  – Full control of supply chain
  – Proprietary security features

• Custom power distribution system

• Workload-specific server optimization
  – Volume drives specialization
Machine Learning Training

Model Complexity
(# of parameters)

- Perceptron 1
- Alexnet 62M
- VGG16 138M
- YOLO, GNMT 210M
- BERT-L 340M
- GPT-2 1.5B
- GPT-3 175B

Example: EC2 P4de

- **ML training monster**
  - 6RU, multi-chassis server
  - Peak power draw 5kW
- **8 PCIe attached NVIDIA A100 GPUs**
  - Massive 826 mm$^2$ with 54B transistors
  - 6,912 cores each – 55,296 cores across all 8
  - 624 TOPS @ INT8 – 4,992 total
- **Mem:** 640GB HBM2e + 1.1TB
- **600 GB/s NVSwitch local**
- **400 Gbps Net with RDMA**
- **8x 1TB NVMe SSDs**
Networking: Mainframe Business Model

- **Central Logic**
  - Custom ASICs
  - Single source

- **Finished H/W Supply**
  - Proprietary
  - Single source

- **System Software Supply**
  - Proprietary
  - Single source

- **Protocol Stack**
  - No provision to alter
  - Single source

---

- **Net Equipment**

- **Central Logic**
  - Commercial ASICs

- **Finished H/W Supply**
  - Custom design
  - Contract manufacture

- **System Software Supply**
  - Linux

- **Application Stack**
  - OSS Protocol Stack

---

- **Commodity Server**

- **Central Logic**
  - Commercial ASICs

- **Finished H/W Supply**
  - Custom design
  - Contract manufacture

- **System Software Supply**
  - Linux

- **Application Stack**
  - OSS Protocol Stack

---

- **Juniper EX 8216**

- **Example**: Juniper EX 8216
  - Fully configured list: $716k w/o optics and $908k with optics for 128 ports

- **Solution**: Merchant silicon, custom H/W, open source protocol/mgmt stack
  - All AWS inside-the-datacenter networking on custom routers

---

http://perspectives.mvdirona.com
Custom ASIC: Nitro

• First AWS ASIC
  – Private server in every server

• Nitro features:
  – Network H/W offload with RDMA
  – Storage H/W acceleration
  – H/W protection & security
  – Hypervisor offload

• Full circle: some mainframe parallels
  – I/O offload to dedicated channel processors
  – RAS & mgmt offload to service processor

• Over 20 million installed
Custom General Purpose CPUs

• Volume supports R&D
  – I’ve long believed Arm could be great server CPUs
    • First blogged in 2009
  – Mobile & IoT volume drive R&D Investment
    • In 2021 Arm crossed 215B processors
• Server Innovation moving to CPU
  – Server innovation moving from board to package
• “AWS Custom H/W” doc review in 2013
  1. Arm will yield a great server processor
  2. Server innovation is moving on-package
Custom Processors

- AWS general-purpose Arm processors
- Graviton 1: re:Invent 2018
- Graviton 2: re:Invent 2019
- Graviton 3: re:Invent 2021
  - 55B transistors
  - 64 Arm Neoverse V1 Cores
  - 64k L1 / 1MB L2
  - 8 DDR5 lanes
  - 7-die multi-chip package
ASIC Acceleration

• Full circle: H/W acceleration again the norm
  – Processing volumes justify specialized H/W
  – Core algorithm stability

• Networking is perhaps the best example
  – All routers have specialized ASICs at core
  – High performance NICs

• Machine learning inference
  – 2018: AWS Inferentia
  – 32 TOPs to 512 TOPS at INT8
  – 100Gbps Networking
ASIC: ML Training

• Training volume drives investment
  – 2021: Trainium

• EC2 Trn1 instance type
  – 16 Trainium ASICs
  – 512GB HBM2 memory
  – 768GB/s intra-instance
  – 800Gbps networking
  – 8TB NVMe storage
  – Stochastic rounding
Nitro SSD

• Problems with commercial SSDs
  – High 9 performance variability
  – Unreasonable profit margins

• Nitro SSD
  – Focus: Cost & performance stability
  – High performance NVMe attach
Power Plants

• Manufacturers often used on-site power plants
• Rarely today
• Full circle: Cloud operators now own power plants
• Amazon at 12,000 MW
  – Worlds largest renewable energy purchaser
Closing

• In early days of HPTS many had control of:
  – CPU semiconductor innovation
  – Server design
  – Storage sub-system
  – System software
  – Application stack

• Move to commodity servers lowered costs
  – But made many innovations impractical

• Cloud computing scale brings it all back
HPTS Comes Full Circle

Slides: mvdirona.com/jrh/work
Email: james@amazon.com
Blog: perspectives.mvdirona.com