Scaling Out High Performance Computing

9th Annual MVAPICH User Group (MUG) Meeting

Luiz DeRose
Director, HPC Cloud Engineering
Oracle
August 25, 2021

Safe harbor statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Outline

Trends in computer architecture

The impact on programming and middleware environments

The impact on applications

“Software is getting slower more rapidly than hardware becomes faster”

Niklaus Wirth 1995

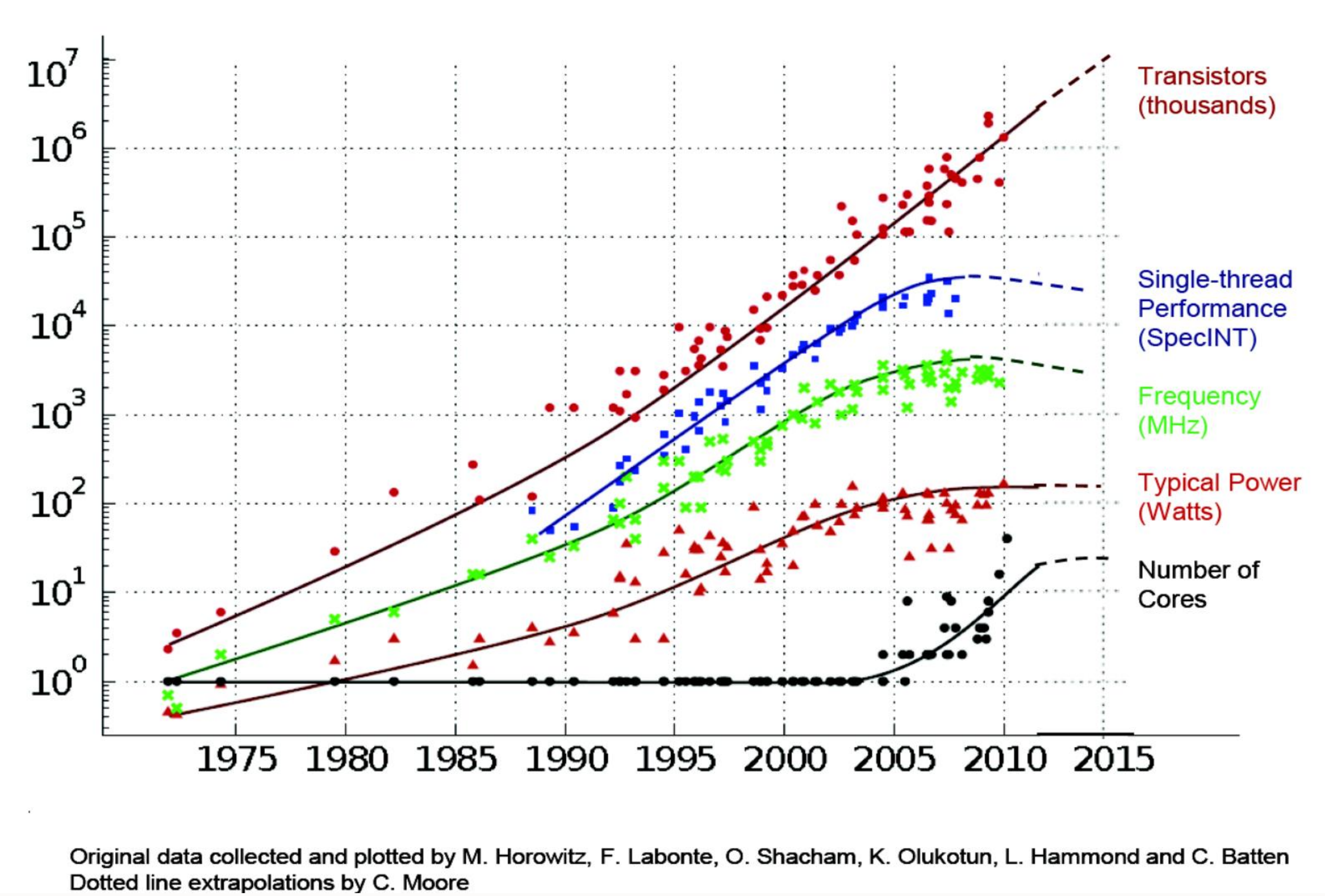
HPC on the cloud

Oracle Cloud Infrastructure

The feel of on-premise in the cloud

Conclusions

35 years of microprocessor trend data



Trends in computer architecture

Moore's Law – "The density of transistors on a chip will double every two years"

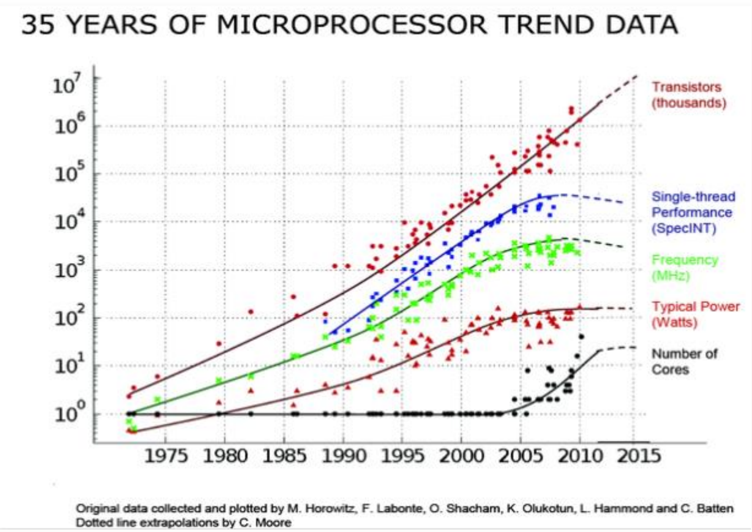
- With smaller chip area the clock cycle was reduced

Dennard's scaling says that performance per watt will also double every two years

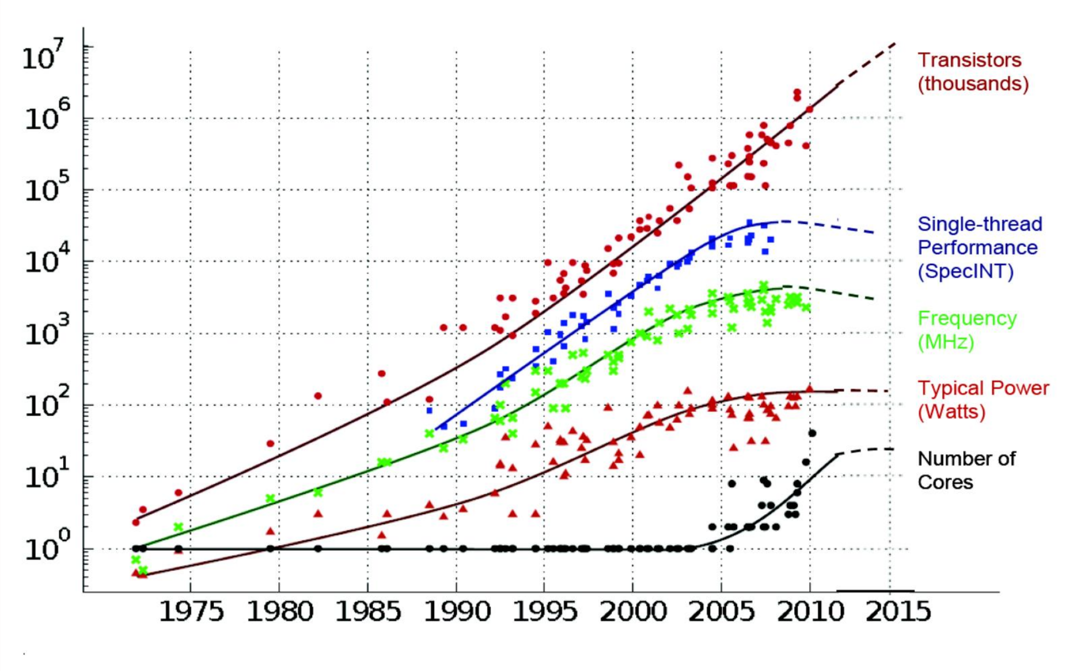
- But due to heat and physics Dennard scaling no longer applied resulting in the clock cycle staying constant and, in some cases, getting higher

Number of transistors on the chip has been increasing, but

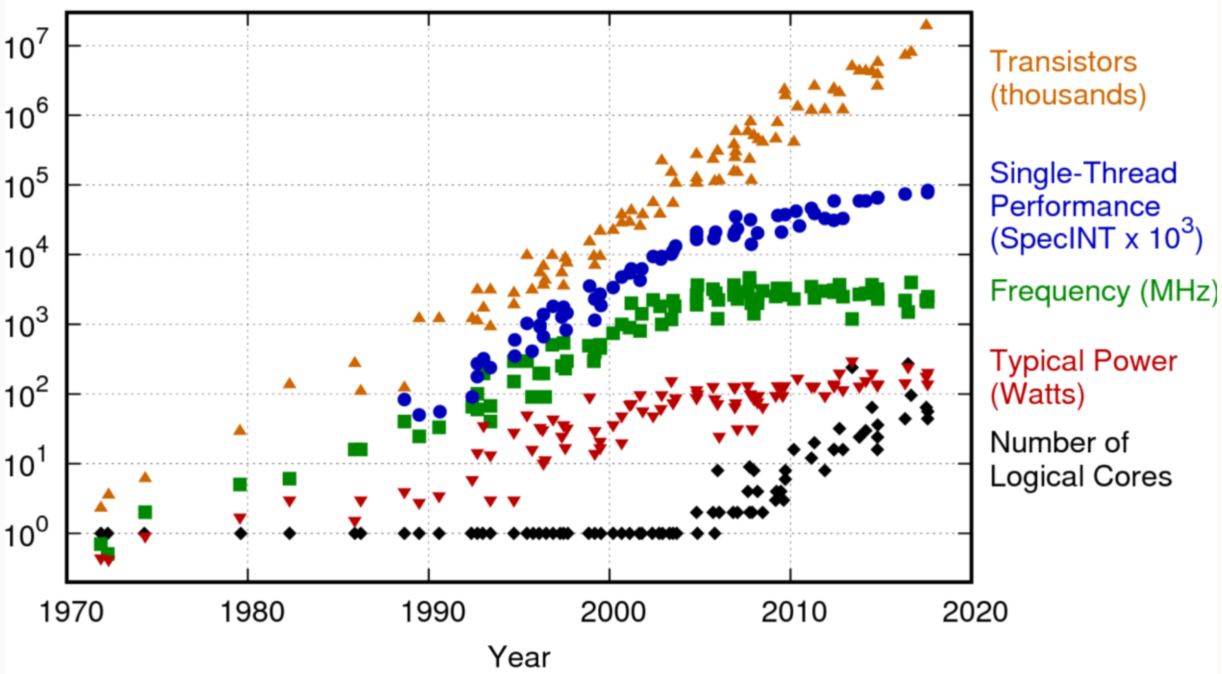
- Ability to increase the transistor count alone, is not enough to sustain further microprocessor development
  - Speed of processor clocks not increasing
    - Too much heat produced
    - Too much power consumed



42 years of microprocessor trend data



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore



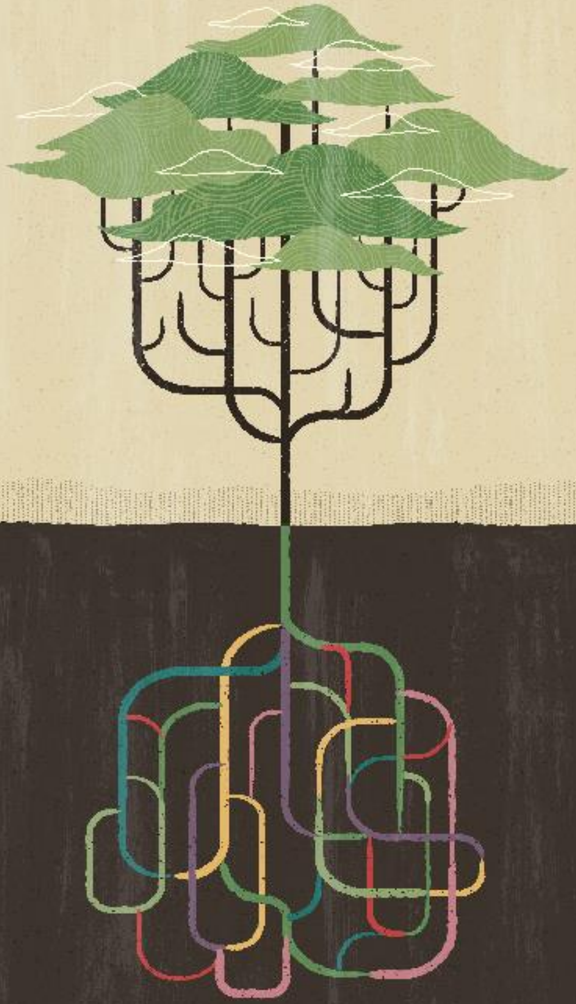
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

Industry response:

multi/many-core designs, GPUs, vectorization

Computer Architecture Trends

Effects on Programming and Middleware Environments

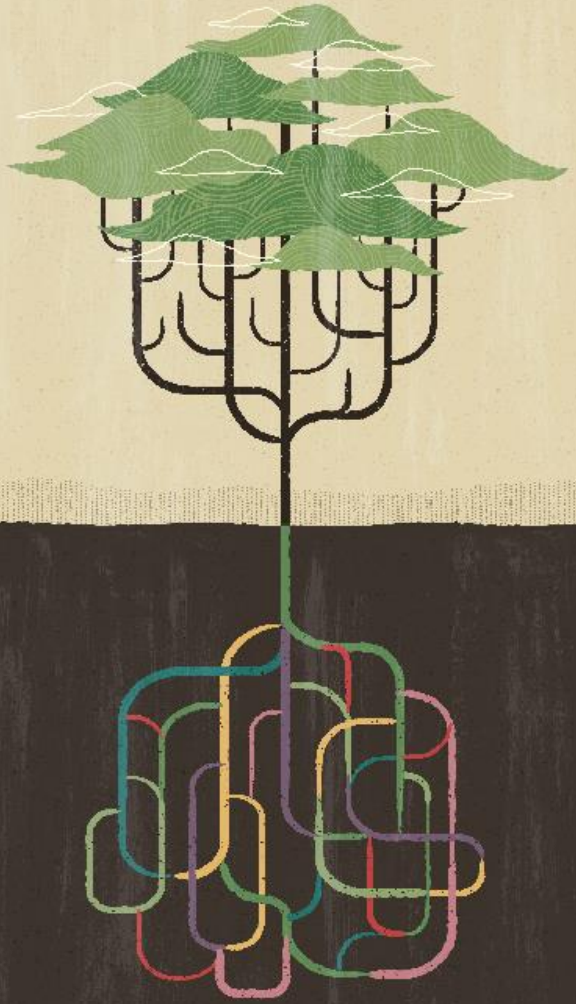


Computer architecture trends
Effects on programming and middleware environments

CPU core vs GPU core

- CPU core: relatively heavy-weight, designed for complex control logic, optimized for sequential programs.
  - Embarrassingly Parallel
  - Tightly-coupled
- GPU core: relatively light-weight, designed with simple control logic, optimized for data-parallel tasks, focusing on throughput of parallel programs.

Computer Architecture Trends



Effects on Applications

Computer architecture trends – effects on applications

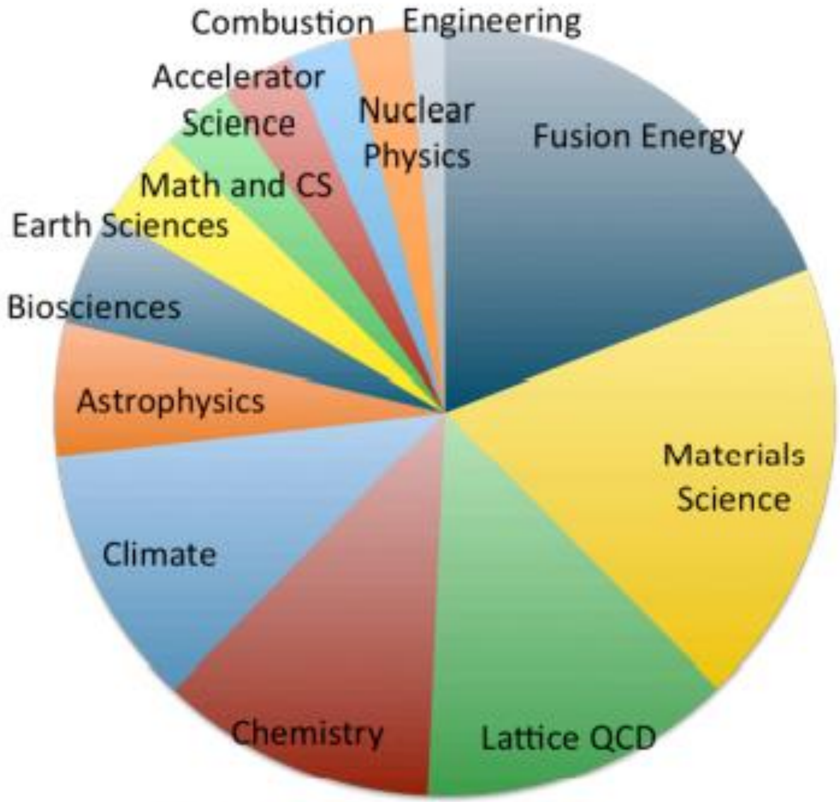
General purpose architectures are optimized for the most widely-used applications

- Mostly commercial and gaming applications

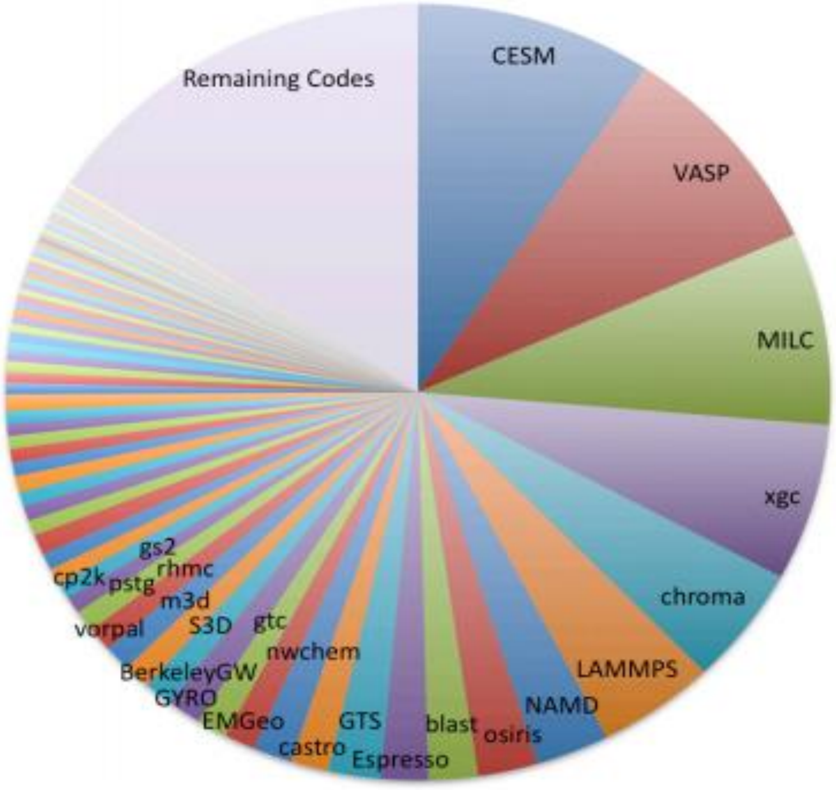
Application requirements are not “one-size-fits-all”

- Some compromises are needed between processor and application development
  - Seldon a one-to-one match between the problem being solved and the processor technology used
  - As applications get more complex, the match is even harder

NERSC workload analysis on Hopper



Hopper CPU hours broken down by science area.



Hopper CPU hours broken down by code usage.

K. Antypas, B.A. Austin, T.L. Butler, R.A. Gerber, C.L. Whitney, N.J. Wright, W. Yang, Z. Zhao
NERSC Tech Report 1163230: https://www.osti.gov/servlets/purl/1163230/ May 2014

Computer architecture trends – effects on HPC applications

Multiple levels of parallelism

- Massively parallel architectures
  - Message passing between nodes

- Wider nodes
  - Nodes are becoming more parallel
    - More processors per node
    - More threads per processor

- Longer vectors and accelerators
  - Vectorization at the lowest level (SSE, AVX, SVE, GPU, …)
  - Vector lengths are getting longer

As nodes get wider, memory bandwidth cannot keep up, resulting in more complex memory hierarchy

On-premises HPC workload challenges

Capacity Planning

- Correctly balancing capacity across planned and unplanned demands
- Supply chain and procurement processes can result in delays of months
- Existing capacity doesn't run our jobs quickly enough

High Costs

- Over-provisioning to ensure burst capacity can lead to overspending
- Cost of lost performance on legacy hardware.
- Power, cooling, networking, storage, hw and sw add to on-premises costs

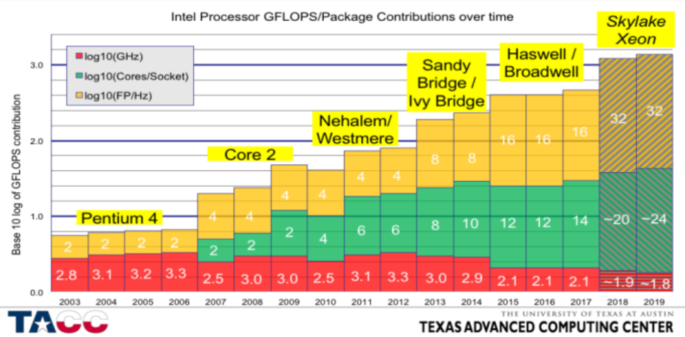
Hard to Optimize Operations

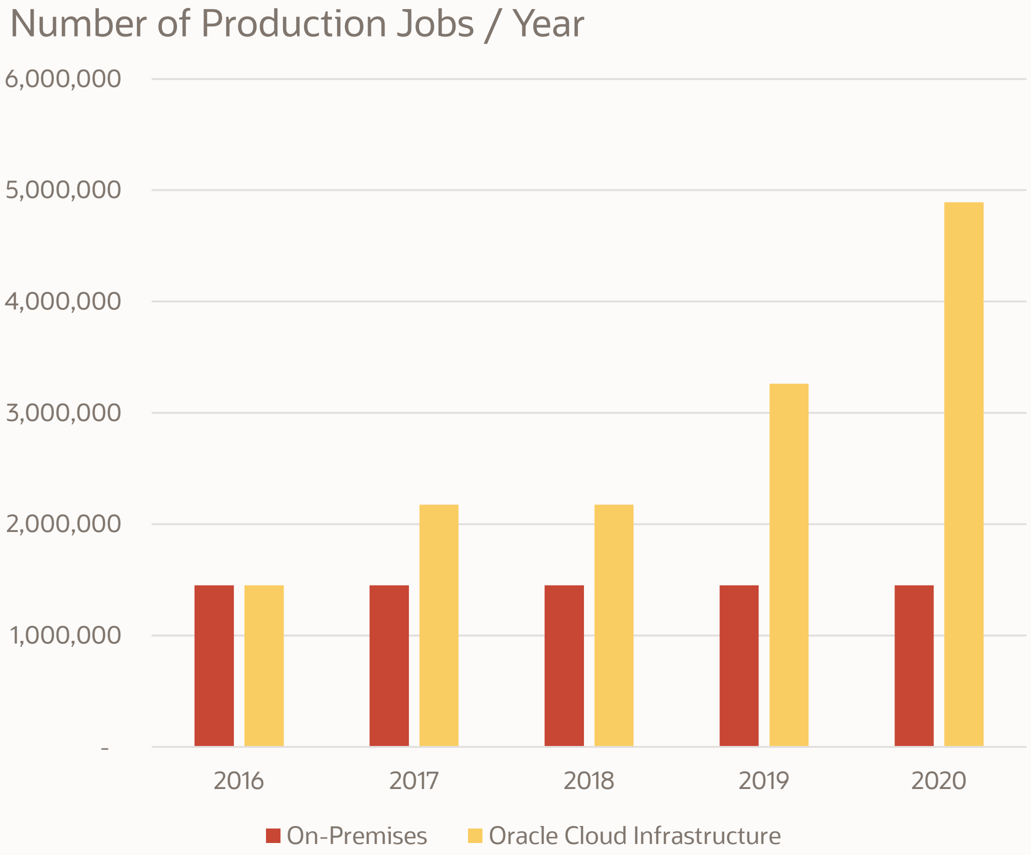
- HPC farms must run at near capacity to maximize ROI
- Under-provisioning can lead to long queuing times for new jobs, productivity
- Staying on top of security and compliance requirements is demanding and missteps can be devastating

On-premises cost of lost performance

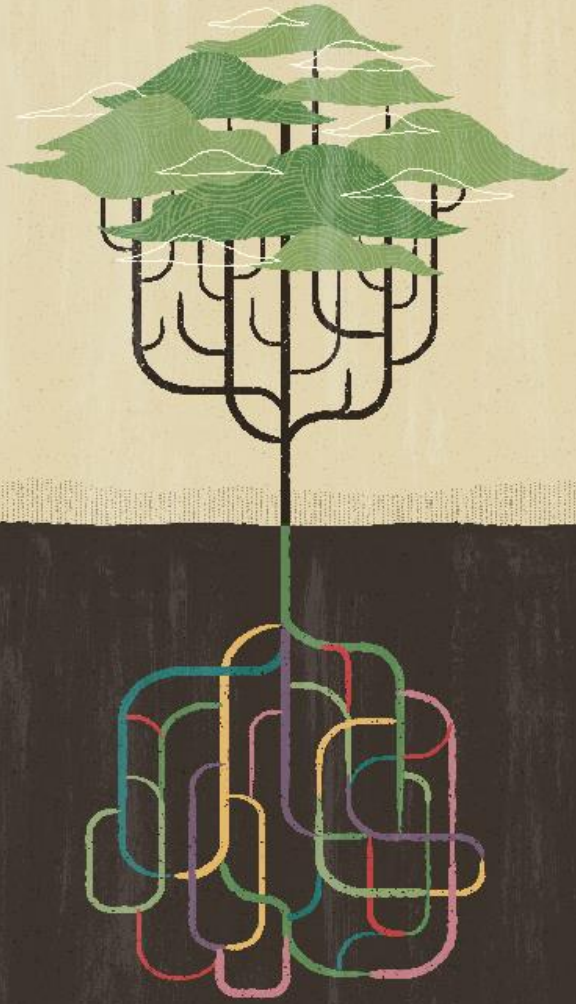
- As a cloud provider we launch new instances running on the latest hardware every year
  - 2016: Intel Broadwell
  - 2017: NVIDIA Pascal
  - 2018: Intel Skylake and NVIDIA Volta
  - 2019: AMD Naples
  - 2020: AMD Rome, NVIDIA Ampere, and Intel Ice Lake

- On-premises deployments lose out in performance gains over the life of the system





HPC in the Cloud



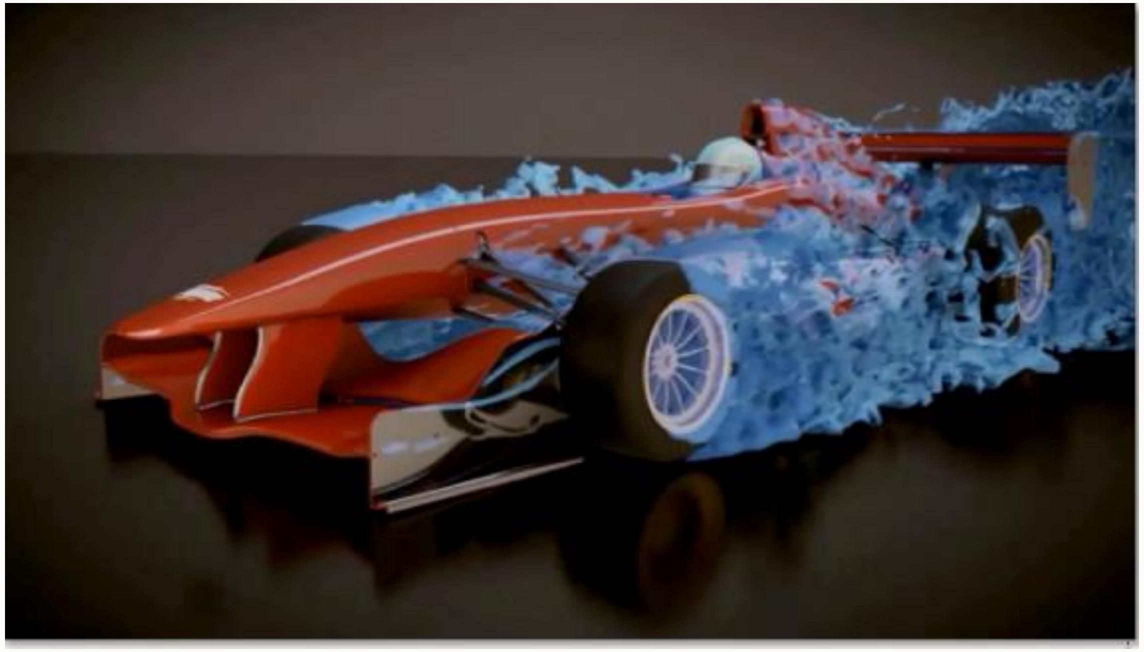
HPC users need an infrastructure that can adapt to their workload needs

HPC is a late moving workload to the cloud

Historically, companies have invested millions of $ in HPC infrastructure on-premise

Initial cloud performance were not great due to virtualization of the OS, network,…

Computer-aided engineering (CAE) and Biosciences have been the early adopters



The HPC cloud market is growing

HPC cloud usage on the rise

HPC market $46.6 billion by 2024

– **Hyperion Research**

Cloud is the fastest growing segment of the HPC market, projected to exceed $3B by 2023

– **Intersect360**

HPC cloud revenue is projected to reach $8.8 billion in 2024, reflecting a 5 years CAGR growth rate (CAGR) of 17.6%

6.9% CAGR for On-prem server

– **Hyperion Research**

HPC investments are paying dividends

An investment in HPC results in an average increase in revenue of $463 per dollar invested in HPC, and an average increase of profits (or cost savings) of $44 dollars per dollar

- **IDC**

HPC cloud vertical forecast

The HPC cloud market will reach about $9 billion in 2024

- Biosciences and CAE have been the early adopters
- Weather, geosciences and academia are showing the highest growth over the forecasted period
- Public sector will have a compound annual growth rate (CAGR) of 20.76%

($M)	2018	2019	2024	2019-2024 CAGR
Biosciences	$778	$1,230	$2,453	14.8%
CAE	$469	$733	$1,540	16.0%
Chemical Engineering	$62	$98	$211	16.6%
DCC & Distribution	$141	$222	$519	18.5%
Economics/Financial	$123	$195	$430	17.1%
EDA	$178	$285	$677	18.9%
Geosciences	$148	$240	$660	22.4%
Mechanical Design	$12	$20	$44	17.1%
Defense	$185	$296	$705	19.0%
Government Lab	$173	$274	$625	17.9%
University/Academic	$123	$197	$528	21.8%
Weather	$26	$42	$220	39.3%
Other	$49	$79	$188	18.9%
Total	$2,467	$3,911	$8,800	17.6%

Hyperion research forecast @ ISC 2021

Why HPC on the cloud is better than on-premises

Run on the latest generation

- CPUs, GPUs, ARM, NVIDIA, Intel, AMD, Oracle Cloud will be running the latest generation hardware
- Eliminate the cost of lost performance
- Eliminate refresh cycles

Pay only for what you use

- Scale your compute cluster for your current workload
- Pay for the correct hardware, at the correct time, at the correct scale
- Use cloud credits for any type of infrastructure or Oracle Cloud service offering

Get out of the DC business

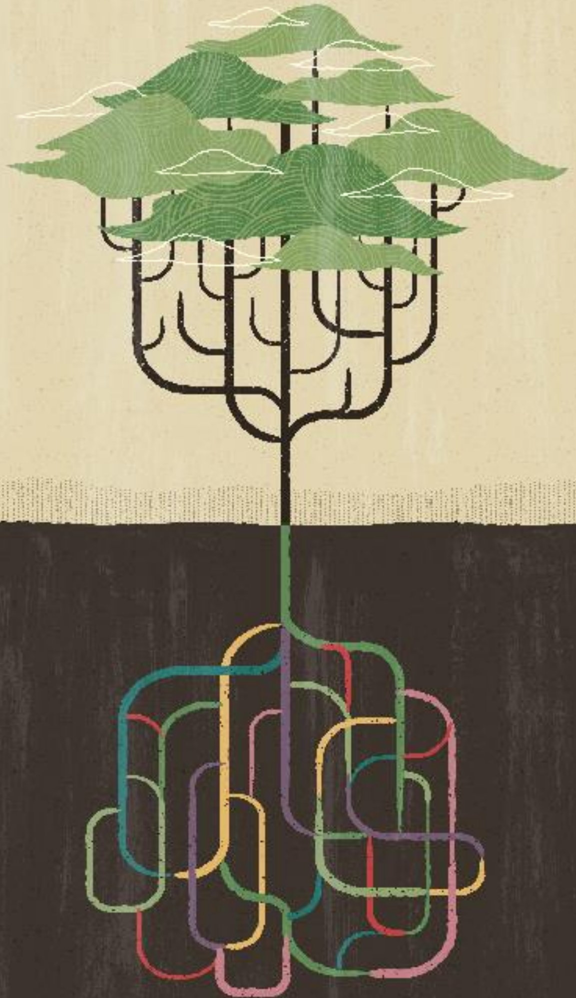
- No more real estate, co-lo's, electricity bills, etc. Focus on what you do best
- Leverage Oracle's public cloud infrastructure, monitoring, maintenance, and management

How users benefit from moving HPC workloads to the cloud

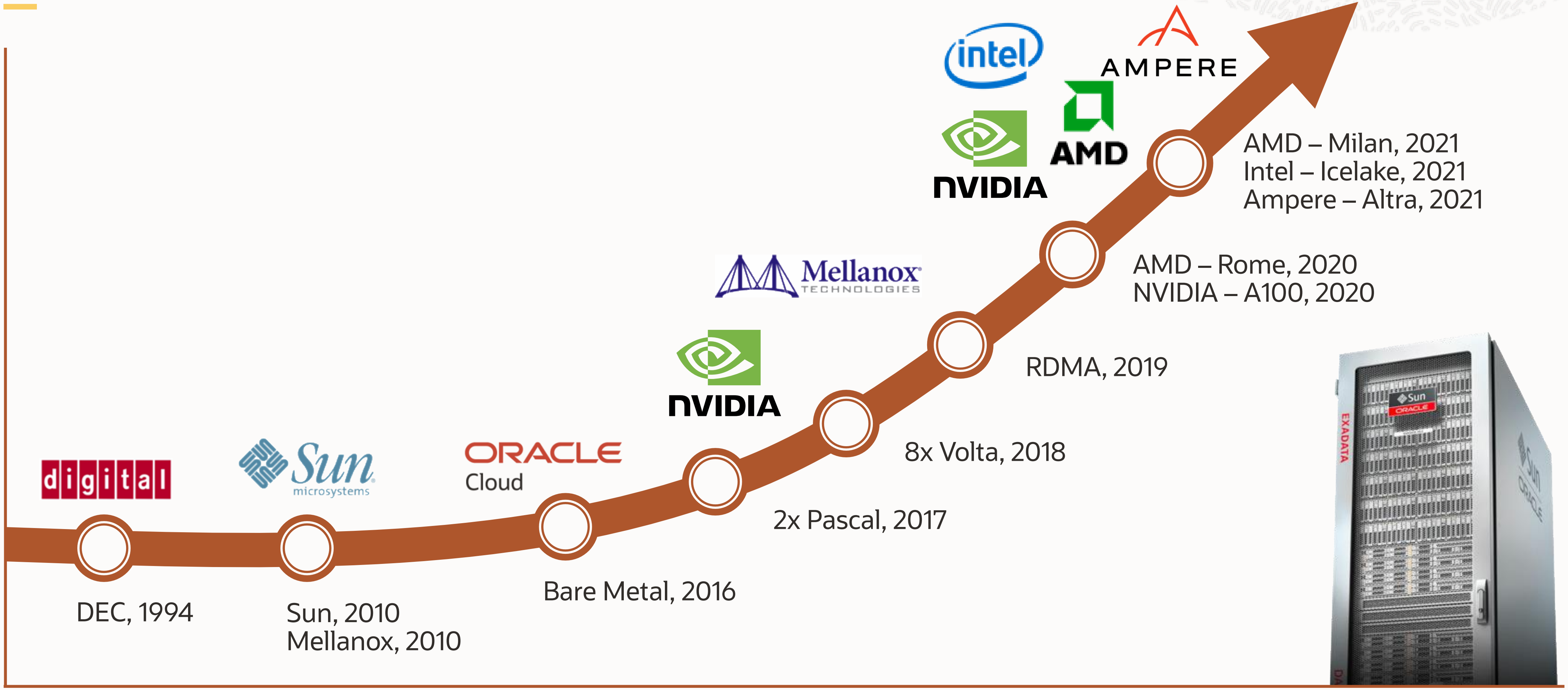
	Cost	Capacity	Flexibility	Performance
Engineering Outcome	Reduce on-premises infrastructure spend and predict costs based on usage	Solve models faster with access to more compute and storage	Run jobs at any time with on demand HPC compute capacity	Best price performance of any cloud vendor. Largest single node machines. Largest available on-node storage
IT Outcome	Eliminate the hardware refresh and capacity planning vicious cycle	Utilize cloud to respond to unplanned jobs and eliminate job queues	Focus on strategic priorities, not managing hardware and software	Next generation performance matches on-premises hpc clusters

Oracle Cloud Infrastructure for HPC

The feel of on-premise in the Cloud



Oracle's HPC journey



The Metal in Bare Metal

“What we really liked about the bare metal offering from Oracle is that there was very little technology between us and the hardware.”

David Standingford
Director and Co-founder

ZENOTECH
SIMULATION UNLIMITED

Why HPC is Better on Oracle Cloud

Bare Metal Compute

- First true bare metal offering
- Better price performance than other cloud providers
- Similar or better performance with on-premises environments

Scalable, Independent Storage

- More NVME local storage than any other provider
- Block storage delivers millions of IOPS at the lowest cost in the industry
- Faster parallel file system performance than any other cloud provider

Specialized Network

- Non-over subscribed, flat 100 Gbps bandwidth
- Only cloud with Network performance SLA
- 100G RDMA networking

Run HPC on leading-edge processors - Latest Compute Hardware

AMD

EPYC Standard Instance

- Based on AMD EPYC architecture
- Both Bare-Metal & VMs
- Cores: 1 – 128 cores (2.25Ghz base with up to 3.4Ghz boost)
- Memory: up to 2TB
- 100Gbps overall network bandwidth

Milan based AMD Instances

- Dense servers with CPU / Memory custom ratios
- BM or VM shapes available via flexible shapes
  - Pick and choose core and memory split based on workload characteristics

intel

Skylake processors

- Intel Gold Xeon
- Cores: 18 cores with 3 GHz Clock speed and 3.7 Max Turbo frequency
- Memory: Up to 768GB
- 6.4TB of Storage per BM

X9 Icelake Intel Instances

- Cores: 1 – 36 cores with High-Frequency (3.6 Ghz Turbo)
- 100G RDMA
- Local NVME SSDs
- Scale to 1000s of cores per cluster

AMPERE

ARM Compute Instances

- Powered by Ampere's Altra Processors
- Cores: BM or VM shapes – very dense, up to 160 cores single threaded performance (3.3Ghz/core)
- Best price performance for general purpose/ massively parallel workloads with up to 30% performance improvement over other x86 compute instances per core

NVIDIA.

GPU Bare-Metal Instances

- NVIDIA Ampere Architecture
- Both Bare-Metal and VMs
- GPUs: up to 8 (for BM)
- Memory: Up to 2 TB
- 1.6 Tbps RDMA
- For intensive applications like genomics, AI deep learning training and inference, data analytics, scientific computing, edge video analytics and 5G services, graphics rendering, cloud gaming…

How about the network

If the MPI tasks in your program are like marathon runners, what should you care most about?

THE FASTEST GUY?

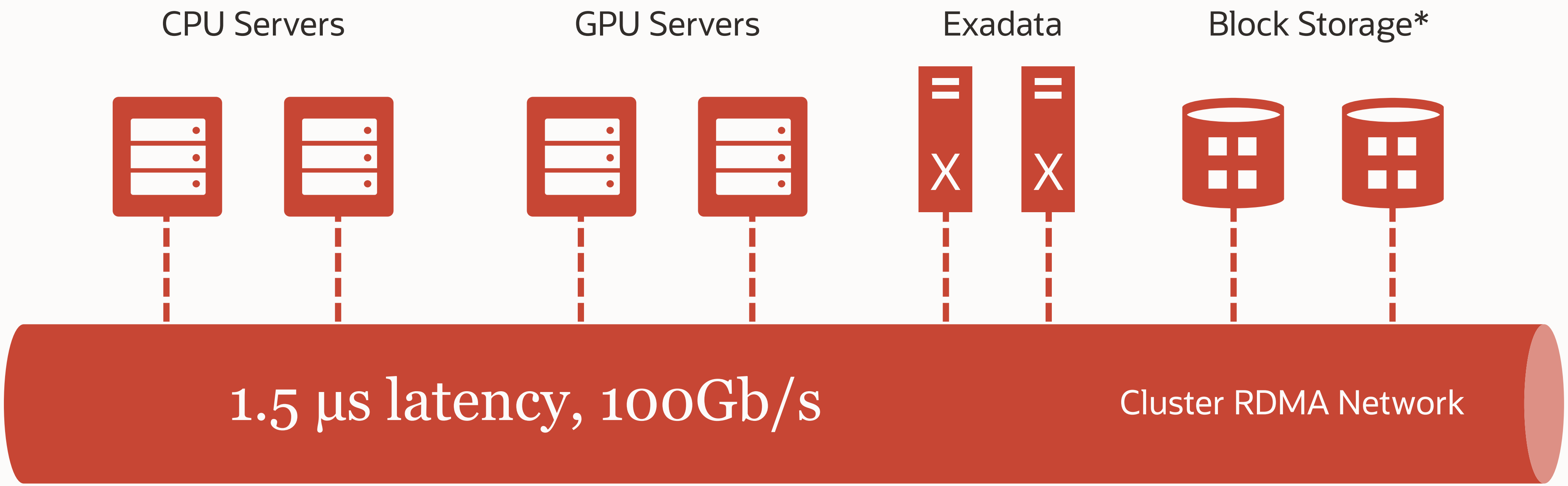
THE AVERAGE?

THE LAST PLACE FINISHER?

Cluster networking

- Low latency
  High Bandwidth
  Predictable performance
  Smart host placement within data center

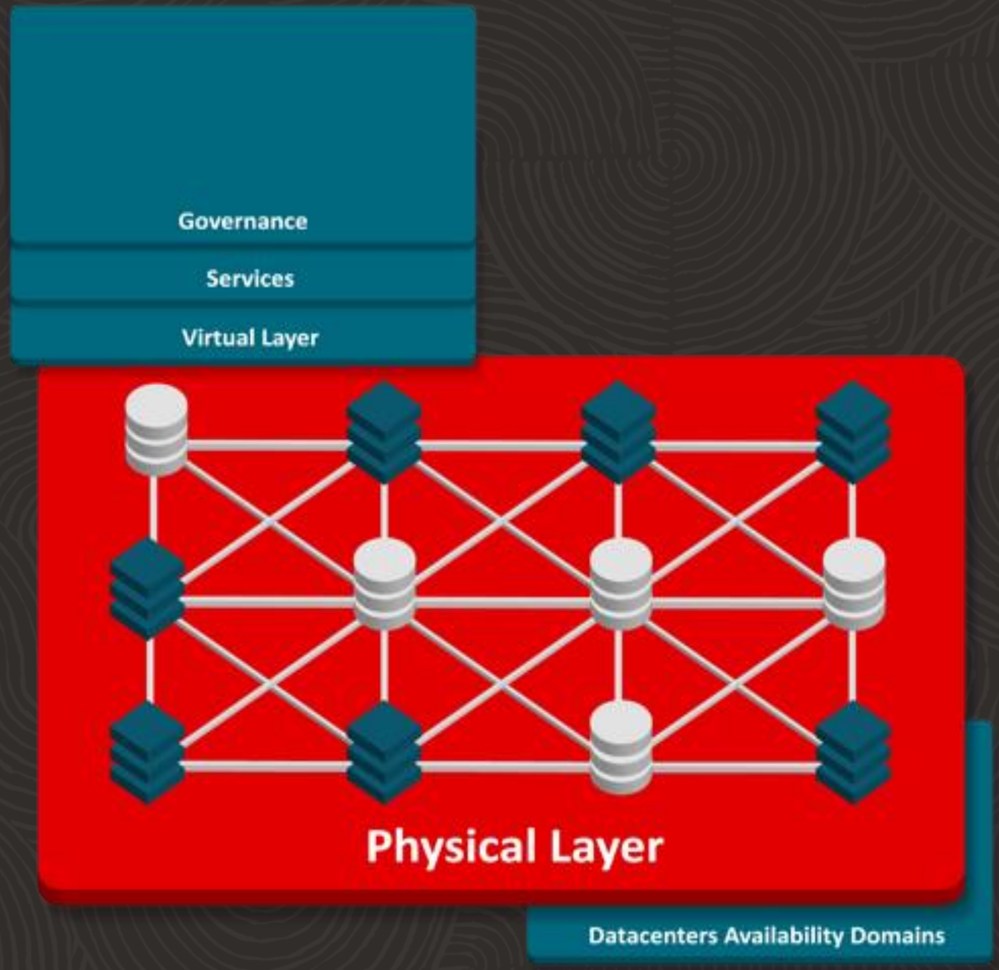
RDMA over converged Ethernet
True hyperscale HPC cloud



For high performance workloads (HPC, Database, Big Data, AI) including the hardest product development workloads like CFD, Crash Simulations, Reservoir Modelling, DNA Sequencing

Fast and Predictable Physical Network Infrastructure

- Non-oversubscribed flat, highly scalable network with ~1 million network ports in each AD
- High speed interconnects: up to 2 x 50Gbps bandwidth
- Predictable, low latency - < 100µs expected one-way latency between hosts in an Availability Domain (AD), <500µs between ADs
- The only cloud network performance SLA



Easy to deploy file systems and high-performance storage

- Deploy HPC parallel file systems on OCI in just 3 clicks
- Choose from a wide set of parallel file systems: IBM Spectrum scale, BeeGFS, Lustre, Quobyte and more
- Achieve 60 -100 GB/sec throughput for HPC parallel file systems
- HPC instance includes 6.4 TB local NVMe
- Block storage volumes up to 1PB

"Oracle Cloud is the only major cloud provider delivering Spectrum Scale in a shared-storage architecture on bare-metal hardware. The Oracle offering has been tuned for the demands of HPC and big-data analytics and offers 10X performance advantage over virtualized cloud offerings, improved flexibility, scalability, and manageability."

Michael Sedlmayer
President, Re-Store

HPC customers cloud utilization

24/7	Burst	Variable Workloads
Turning cluster on 24/7. Using this as a hassle-free datacenter	Hybrid mode, base workload are running on prem and what is overflowing comes to the cloud.	Scheduler is turning machines on/off depending in the queue size.

Support for hybrid cloud

Oracle Public Regions

- Hyperscale cloud regions in 30 worldwide locations

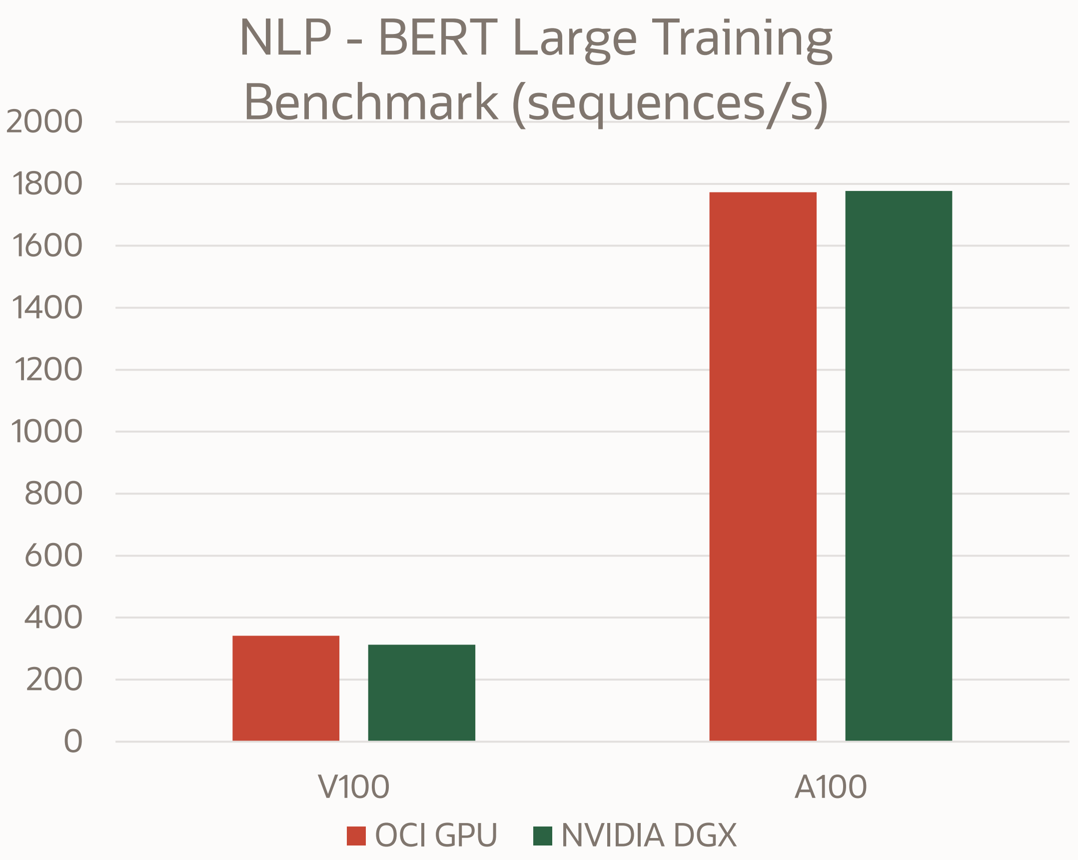
Dedicated Regions

- All OCI services, running in customer data centers

Scaling out - Autoscaling

- Multiple queues with multiple shapes per queue
- Support for all shapes
- Supports both persistent capacity and bursting

On-premise Performance in the Cloud



Cloud GPU shape performance matches on-premise hardware

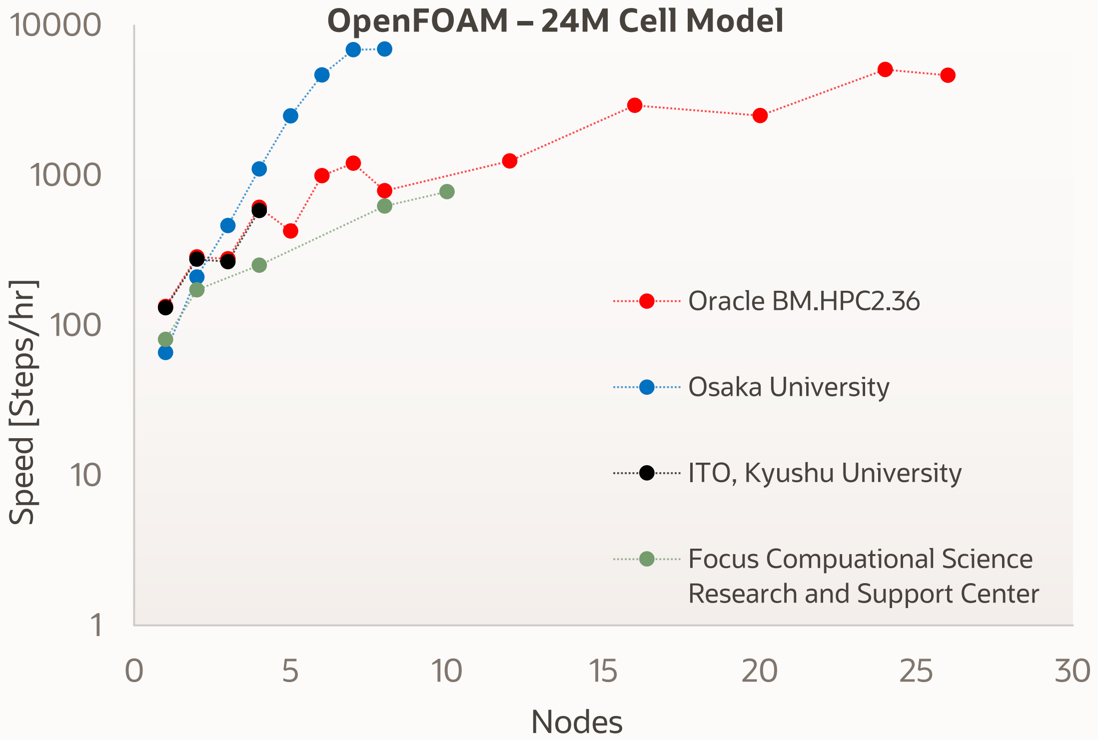
6x uplift on BERT-Large training performance

Improved price performance upgrading to the latest generation

V100 2.95/gpu-hour
A100 3.05/gpu-hour

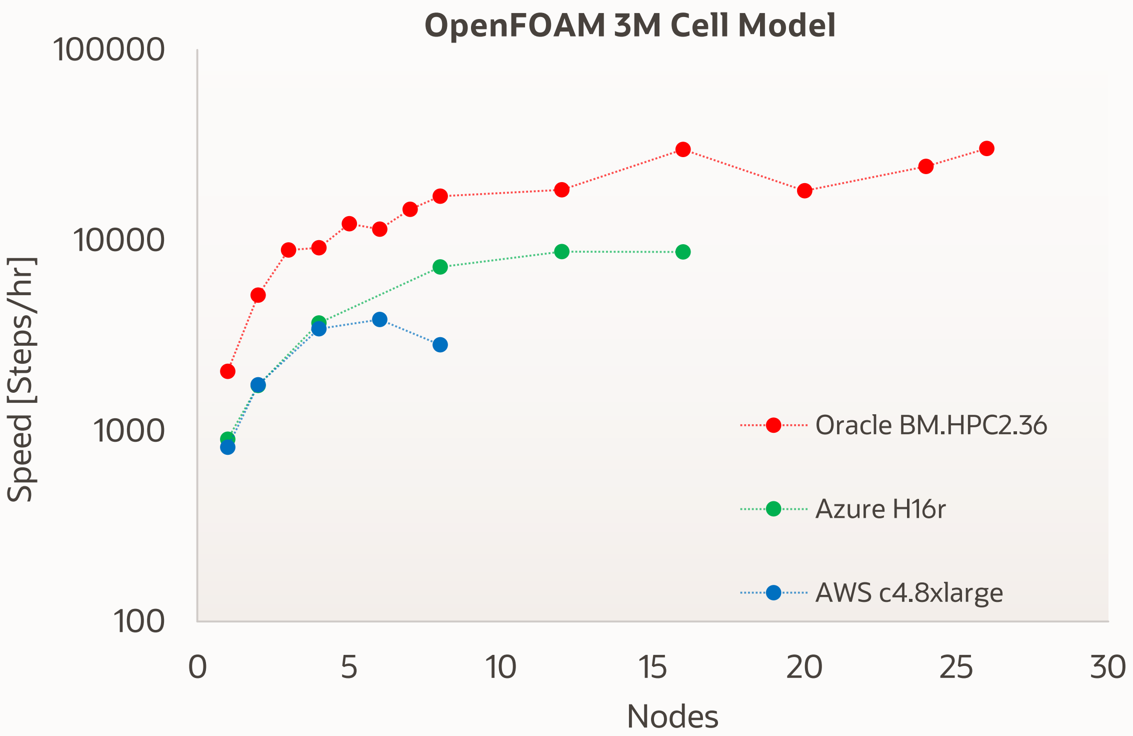
OCI offers supercomputer-level performance

OCI provides on-premises levels of performance



OCI HPC vs On-premises Supercomputers

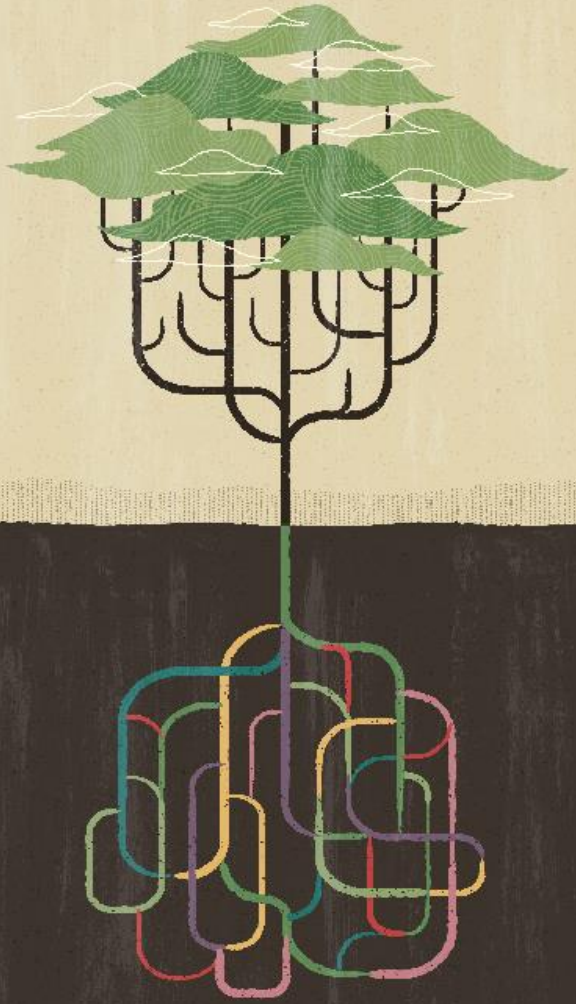
OCI outperforms AWS by 164% and Azure by 278%



OCI vs. competition

Evaluation performed by Open CAE

Conclusions



Conclusions

The trends in microprocessor technologies are influencing HPC significantly

We no longer have a dominant processor architecture

- Application requirements are not “one-size-fits-all”
- Architecture diversity is exciting for research, but
  - Not so much for selection of on-premise system
  - It is also an issue for system software developers

HPC on the cloud is growing significantly

Oracle Cloud Infrastructure offers supercomputer-level performance

- An infrastructure that can adapt to the workload needs

ORACLE

RUN HPC

Thank You