The HPC Ecosystem at AWS

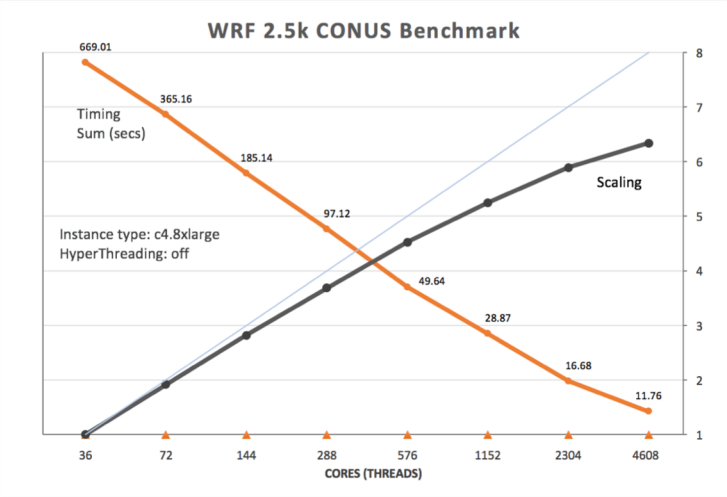
HPC and the Cloud? Surely you must be joking.

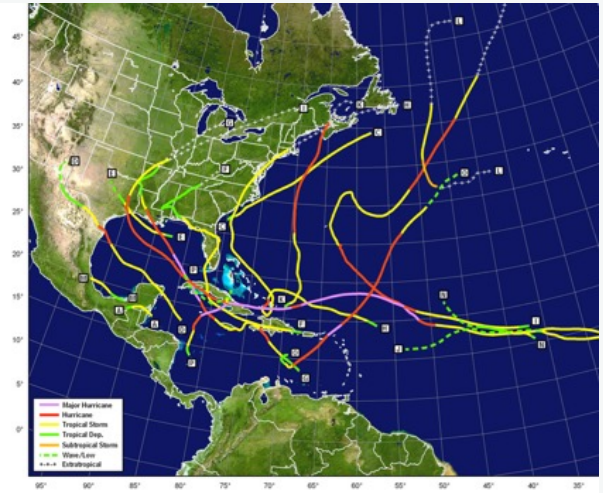
Brian W Barrett

Principal Engineer, HPC Engineering, AWS

HPC in AWS Examples

Tightly-coupled HPC – weather





HPC in aerospace

Boom leverages Rescale and AWS to enable supersonic travel

- Simulated vortex lift with 200M cell models on 512+ cores
- Increased simulation throughput: 100 jobs in parallel with 6x speedup per job → 600x speedup
- Eliminated IT overhead, including server capital costs & in-house IT and software teams
- Elastic HPC capacity and pay-as-you-go AWS clusters allow business agility & ability to scale

- "Rescale's ScaleX cloud platform is a game-changer for engineering. It gives Boom computing resources comparable to building a large on-premise HPC center. Rescale lets us move fast with minimal capital spending and resources overhead."

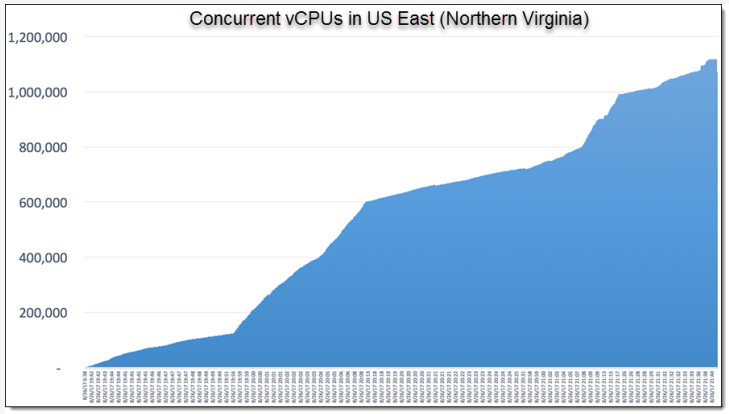
- Josh Krall
- CTO & Co-Founder

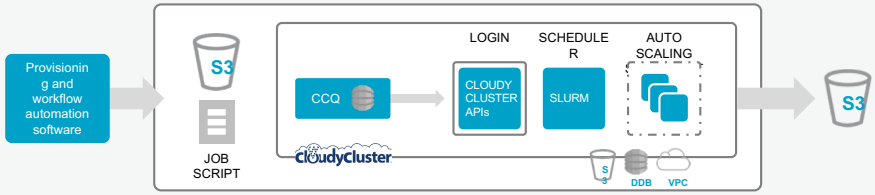
1.1M vCPUs for Machine Learning

The graph highlights the elastic, automatic expansion of resources.

Clemson took advantage of the new per-second billing for EC2 instances.

The vCPU count usage is comparable to the core count on the largest supercomputers in the world.





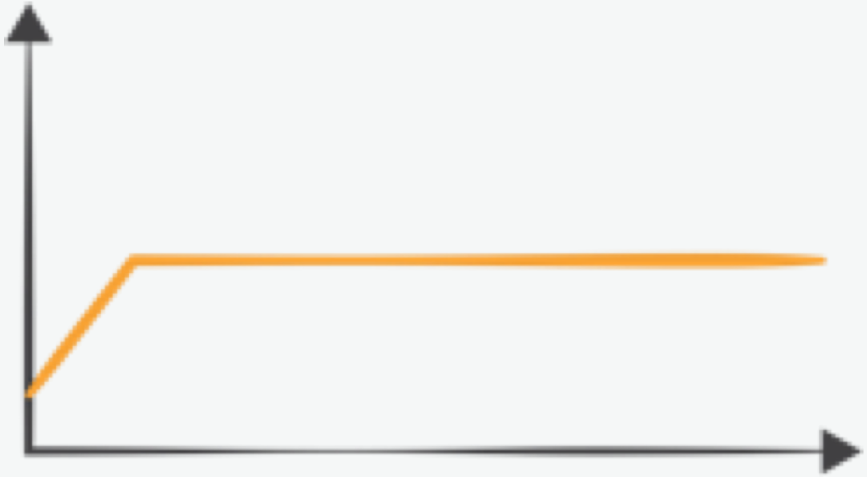
A group of researchers from Clemson University achieved a remarkable milestone while studying topic modeling, an important component of machine learning associated with natural language processing, breaking the record for creating the largest high-performance cluster in the cloud by using more than 1,100,000 vCPUs on Amazon EC2 Spot Instances running in a single AWS region.

What Customers Want

Customer Asks

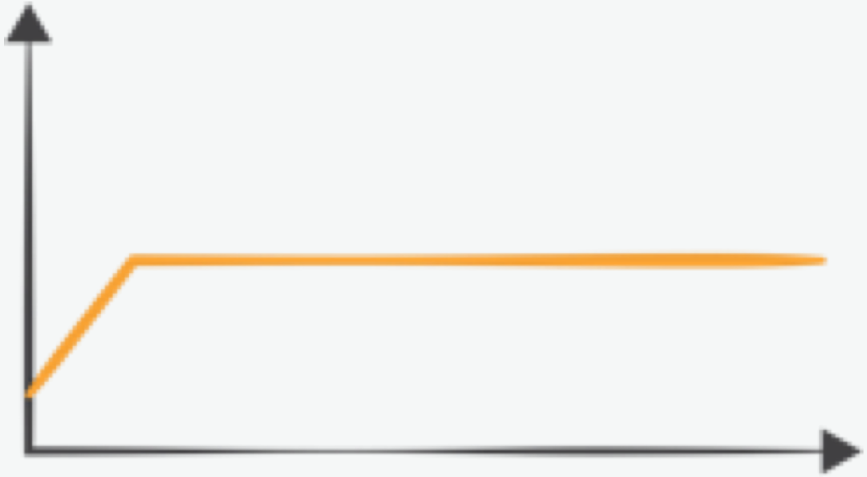
- Elasticity
- Time to completion
- Flexibility
- TCO/ROI

Cluster Load



Ideal

Cluster Load



Ideal

Reality

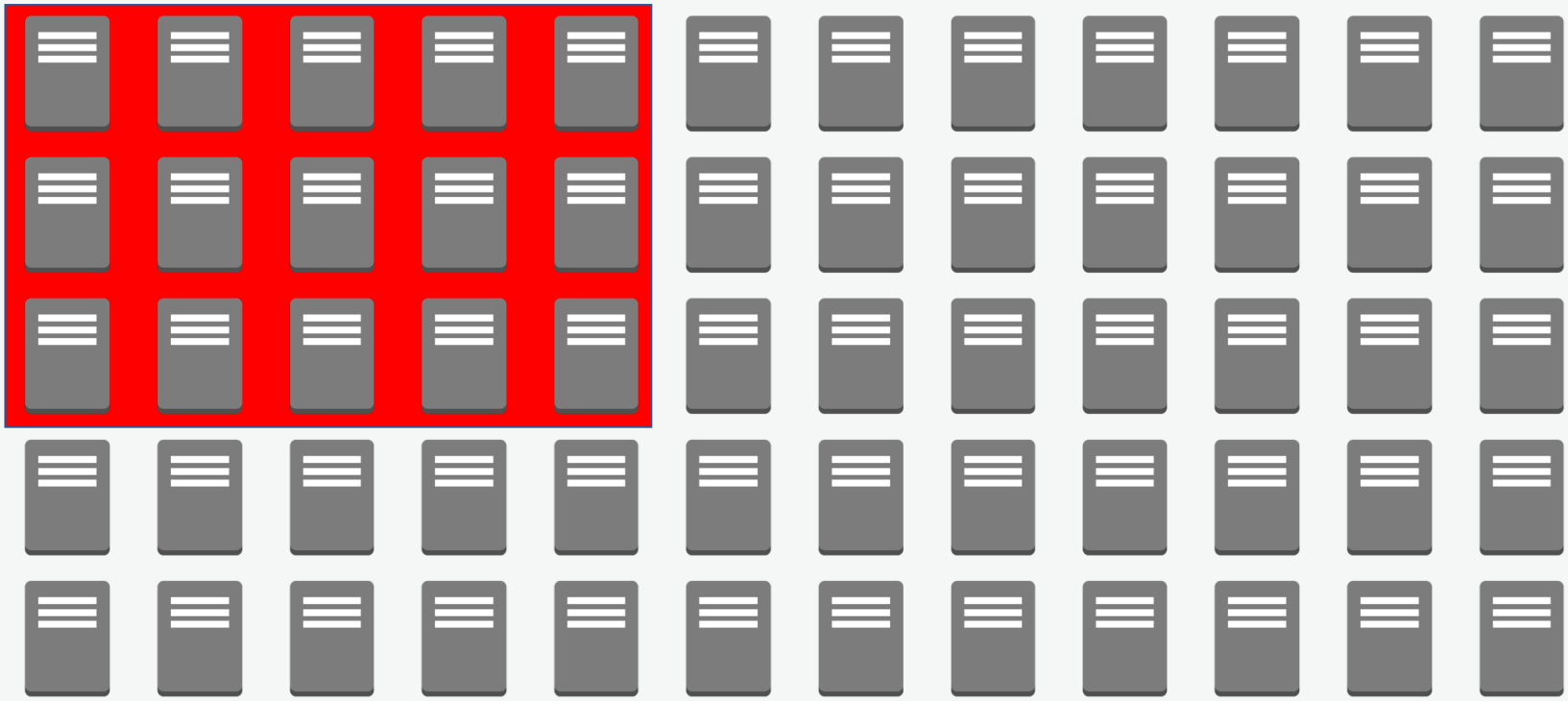
Review: C5 Instances – “Compute”

- Custom Intel Xeon Scalable processor (Skylake) @ 3.0 GHz
  - Turbo up to 3.5 ghz
- Supports AVX512
- C-state controls

Model	Cores	Memory (GiB)	Network Performance
c5.large	1 (2 vCPU)	4	Up to 10Gbps
c5.xlarge	2 (4 vCPU)	8	Up to 10Gbps
c5.2xlarge	4 (8 vCPU)	16	Up to 10Gbps
c5.4xlarge	8 (16 vCPU)	32	Up to 10Gbps
c5.9xlarge	18 (36 vCPU)	72	10Gbps
c5.18.xlarge	36 (72 vCPU)	144	25Gbps

We Confess... We Like Ethernet..

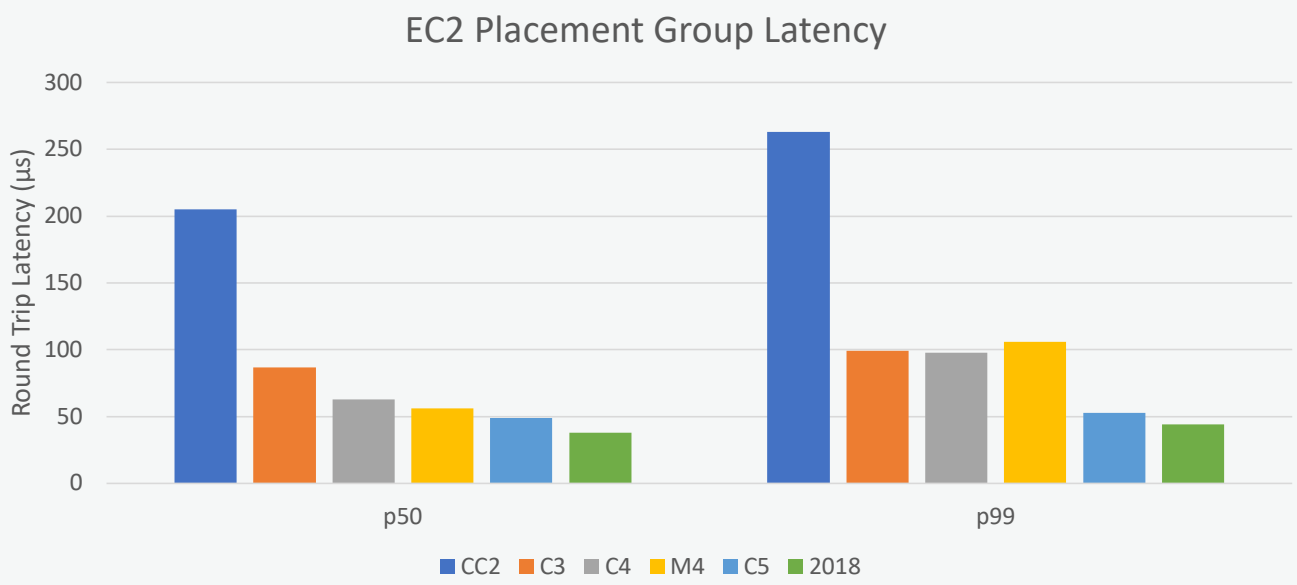
But Ethernet?



But Ethernet?

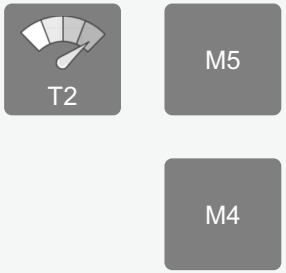
Ok, so Latency?

- We can add bandwidth; it'll cost money, but it's technically easy
- Latency is harder…



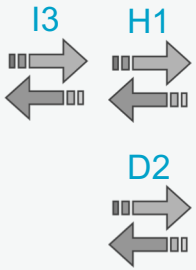
AWS compute instance types

General Purpose

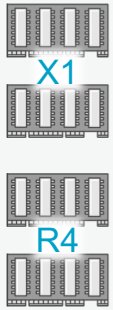


Compute Optimized

Storage and I/O Optimized

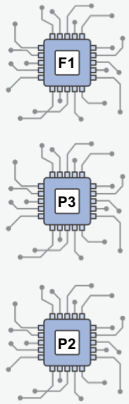


Memory Optimized

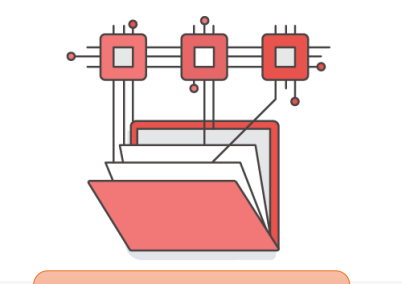


GPU Graphics

GPU and FPGA Compute

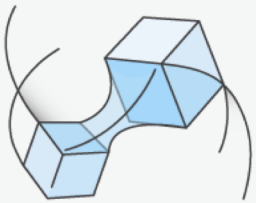


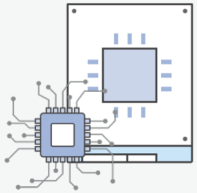
AWS Storage is a Platform



Amazon EFS

File

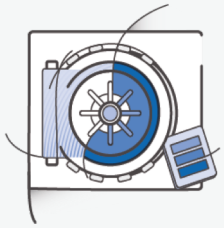




Amazon EBS

Amazon EC2 Instance Store

Block

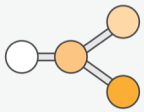


Amazon S3 / S3-IA

Amazon Glacier

Object

Data Transfer



Internet/ VPN

AWS Direct Connect

Amazon CloudFront

S3 Transfer Acceleration

ISV Connectors

Storage Gateway

AWS Snowball

Amazon Kinesis Firehose

So About that Custom Hardware?

Slightly Sillier Example…

Ruggedized case
“8.5G impact”

80 TB
10 GE network

E-ink shipping label

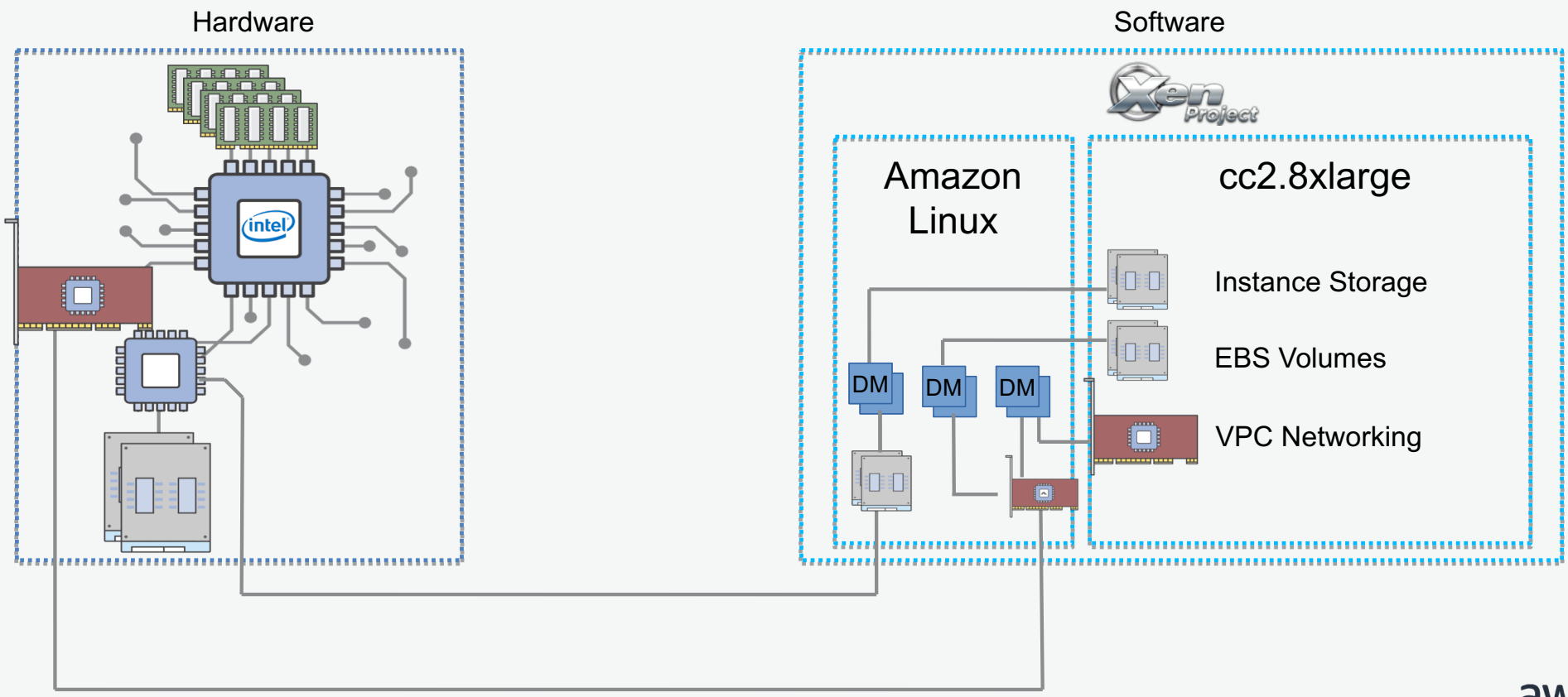
Rain- and dust-resistant

Tamper-resistant case and electronics

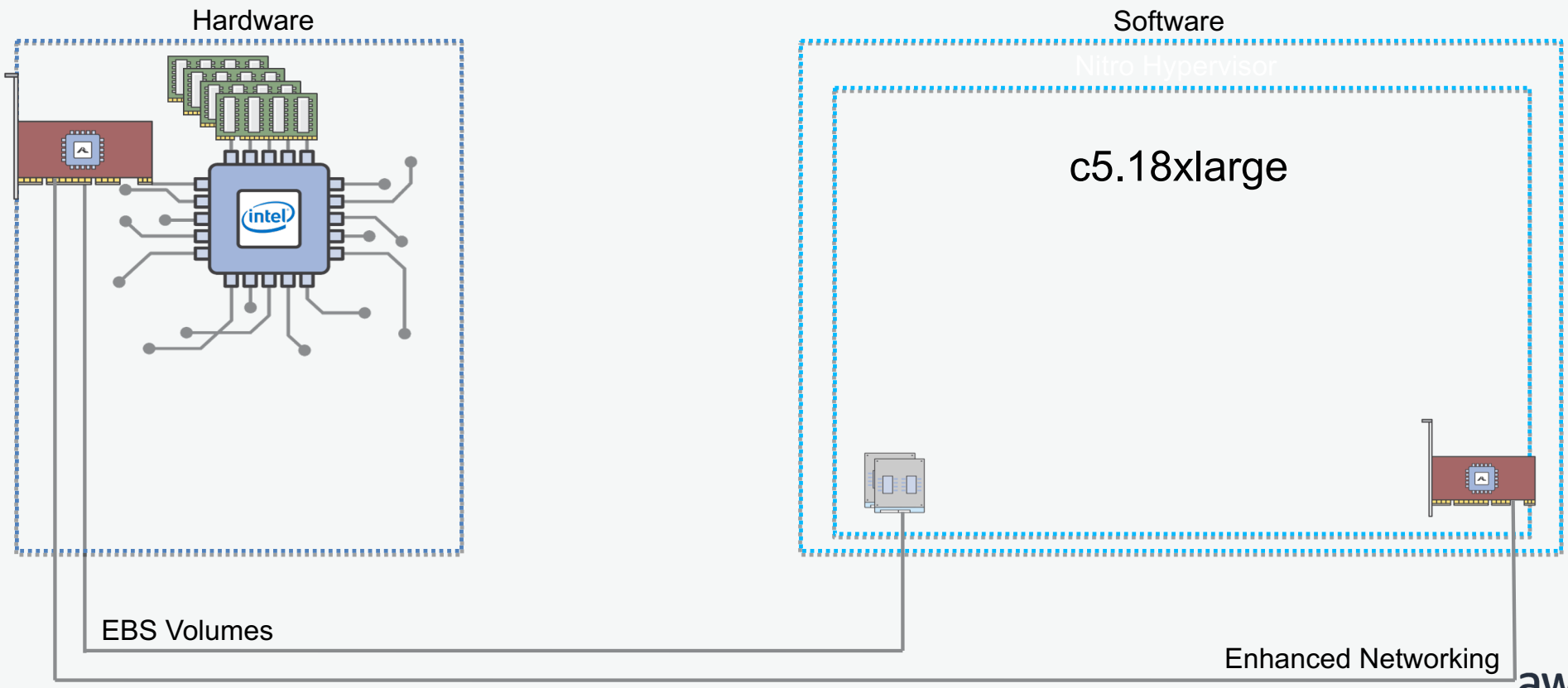
All data encrypted end-to-end

aws

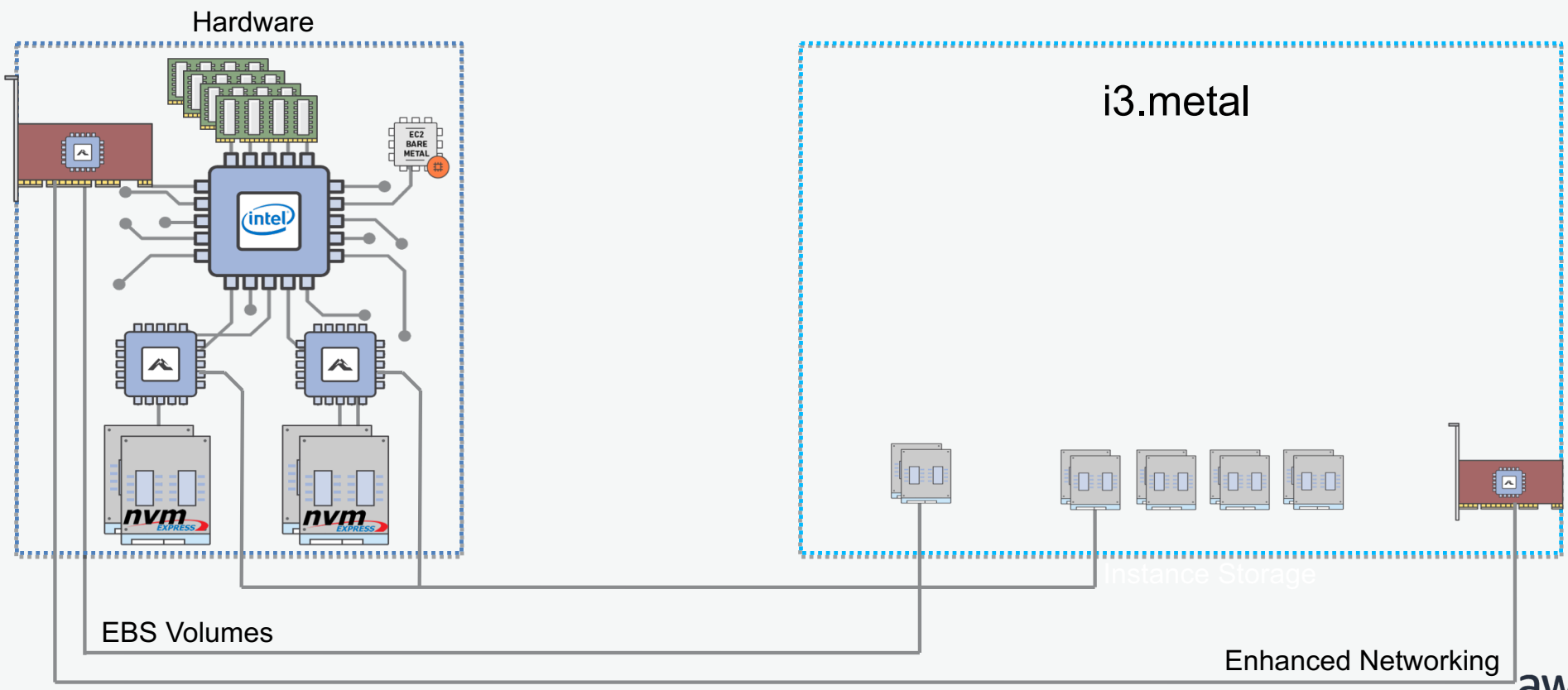
What You Think AWS Looks Like



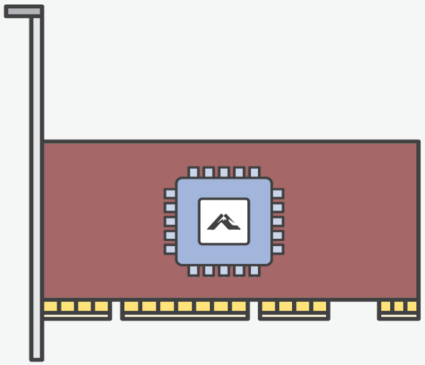
What AWS Looks Like Today

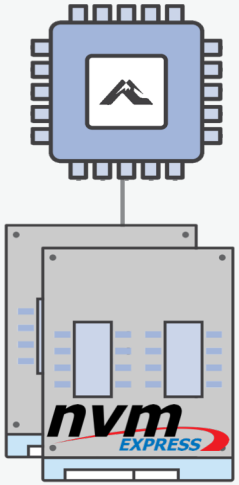


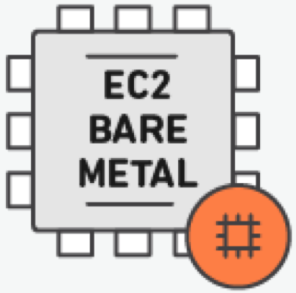
Or if you prefer not to have a hypervisor



Where we're innovating







FPGA Ecosystem

- Building an ecosystem important
- FPGA instance type, 8 FPGAs per instance
  - 16 nm Xilinx UltraScale Plus FPGA
  - 64 GiB DDR4 ECC protected memory, with a dedicated PCIe x16 connection.
  - Each FPGA contains approximately 2.5 million logic elements and approximately 6,800 Digital Signal Processing (DSP) engines.
- Reusable Amazon FPGA (AFI) to build marketplace of FPGA options
- Developer AMI to build your own
  - prepackaged tool development environment
  - scripts and tools for simulating your FPGA design, compiling code, and building and registering your AFI.

Thank you!

Brian Barrett

bbarrett@amazon.com

https://aws.amazon.com/hpc/

Thanks to Adam Boeglin , Anthony Liguori, Alex Maushevsky, and David Pellerin and for slides and data

aws