HPC on AWS
Service Options & Hardware Choices

Matt Koop
Principal Solutions Architect, Compute & HPC
mkoop@amazon.com

MVAPICH Users Group
August 24, 2021
Key services and hardware that enable HPC on AWS

- **Compute**
  - Amazon EC2
  - Bare metal performance
  - AWS Graviton 2
  - EC2 Ultra Clusters

- **Networking**
  - Elastic Fabric Adapter (EFA)

- **Storage**
  - Amazon FSx for Lustre

- **Workflow**
  - MVAPICH2-X-AWS is designed for EFA
  - AWS ParallelCluster
  - AWS Batch
## EC2: Broad and Deep Instance Choice

### Categories
- General purpose
- Burstable
- Compute intensive
- Memory intensive
- Storage (high I/O, dense)
- GPU compute
- Graphics intensive

### Capabilities
- Choice of processor (AWS, Intel, AMD)
- Fast processors (up to 4.5 GHz)
- High memory footprint (up to 24 TiB)
- Instance storage (HDD and SSD)
- Accelerated computing (GPU, FPGA, and ASIC)
  - Networking (up to 400 Gbps)
  - Bare metal
  - Size (Nano to 32xlarge)

### Options
- Linux, Unix, Windows, macOS
- Amazon EBS
- Amazon Elastic Inference
- Elastic Fabric Adapter

### 400+ Instance Types
for virtually every workload and business need
EC2 “Instance” host architecture

- Server
  - 70%
  - 30%
- Customer Instances
- Hypervisor
  - Networking
  - Storage
  - Management, Security, and Monitoring
2012 EC2 “Instance” host architecture

- Customer Instances
- Hypervisor
- Storage
- Management, Security, and Monitoring
- Networking

Server:
- 70%
- 30%
2013 EC2 “Instance” host architecture

- 70% of the resources are allocated to Customer Instances.
- 30% of the resources are allocated to Hypervisor.
- The remaining 10% are allocated to Management, Security, and Monitoring.

Networking and Storage components are also indicated as integral parts of the system.
The Nitro Architecture

- **Server**
  - ~100%

- **Customer Instances**

- **Hypervisor**

- **Nitro**
  - Networking
  - Storage
  - Management, Security, and Monitoring
Metal vs. Nitro Hypervisor Instances on AWS

- AWS offers “.metal” instances, which remove the hypervisor entirely
- The Nitro Hypervisor has minimal overhead in the evaluated HPC applications/benchmarks

AWS Graviton processors

- Custom AWS silicon with 64-bit Arm Neoverse cores
- Targeted optimizations for cloud-native workloads
- Rapidly innovate, build, and iterate on behalf of customers
AWS Graviton 2

- 64 Arm® Neoverse™ N1 cores
- Arm v8.2 compliant
- Worked closely with Arm on creation of N1
  - Large 64KB L1 caches and 1MB L2 cache per vCPU
  - Coherent Instruction cache
  - Lower overheads of interrupts, virtualization, and context switching
  - 4-wide front-end with 8-wide dispatch per issue
  - Dual-SIMD units
  - Data types to accelerate ML inference: int8 and fp16
- Every vCPU is a physical core
  - No simultaneous multithreading (SMT)
- No NUMA concerns
EC2 UltraClusters of P4d instances

On-demand access to a world-class supercomputer

Over 4,000 A100 GPUs

Fully non-blocking petabit-scale network infrastructure

High-throughput, low-latency storage from Amazon FSx for Lustre
Key services and hardware that enable HPC on AWS

- **Compute**: Amazon EC2
  - Bare metal performance
  - AWS Graviton 2
  - EC2 Ultra Clusters
- **Networking**: Elastic Fabric Adapter (EFA)
  - MVAPICH2-X-AWS is designed for EFA
- **Storage**: Amazon FSx for Lustre
- **Workflow**: AWS ParallelCluster
  - AWS Batch
Elastic Fabric Adapter (EFA)

**Elastic Fabric Adapter**
- OS bypass
- GPUdirect and RDMA
- Libfabric core supports wide array of MPIs and NCCL

**Scalable Reliable Datagram**
- ECMP-enabled packet spraying
- Cloud-scale congestion control
- Fast recovery from packet loss or link failure

EFA Software Stack

- **MVAPICH2-X-AWS** is implemented on rdma-core (verbs)
  - Re-ordering with copy-out
  - Immediate data for sequence IDs
  - Packetization for large messages

- Open MPI, Intel MPI, and NCCL-OFI are implemented using libfabric.
- The EFA libfabric provider implements ordering, packetization, and additional semantic options

*S. Chakraborty, S. Xu, H. Subramoni and D. K. Panda, Designing Scalable and High-Performance MPI Libraries on Amazon Elastic Adapter, Hot Interconnect, 2019*
## Wide Variety of Instance Types Available with EFA

### Graviton2
- **C6gn.16xlarge**
  - 1S, 64c, 2GB/core
  - 100Gb

### Storage-Dense
- **i3en.{24xlarge,metal}**
  - 60 TB NVMe
  - 100Gb

### x86
- **“Ice Lake” m6i.32xlarge**
  - 2Sx 32c, 8GB/core
  - 50Gb
- **“Cascade Lake” m5n.{24xlarge,metal}**
  - 2Sx 24c, 8GB/core
  - 100Gb
- **“Cascade Lake” m5zn.24xlarge (4.5GHz)**
  - 2Sx 12c, 8GB/core
  - 100Gb
- **“Cascade Lake” r5n.{24xlarge,metal}**
  - 25x 24c, 8GB/core
  - 100Gb
- **“Skylake” c5n.{18xlarge,metal}**
  - 25x 18c, ~4GB/core
  - 100Gb

### Accelerator
- **p4d.24xlarge**
  - 8x A100 GPUs
  - 400Gb
- **p3dn.24xlarge**
  - 8x V100 GPUs
  - 100Gb
- **inf1.24xlarge**
  - 16 Inferentia chips
  - 100Gb
- **G4dn.{16xlarge,metal}**
  - 8x T4 GPUs
  - 100Gb
Key services and hardware that enable HPC on AWS

- **Compute**
  - Amazon EC2
  - Bare metal performance
    - AWS Graviton 2
    - EC2 Ultra Clusters

- **Networking**
  - Elastic Fabric Adapter (EFA)

- **Storage**
  - Amazon FSx for Lustre
  - MVAPICH2-X-AWS is designed for EFA

- **Workflow**
  - AWS ParallelCluster
  - AWS Batch
Amazon FSx for Lustre

Parallel file system

- High and scalable performance
- +100 GiB/s throughput
- Millions of IOPS
- Consistent submillisecond latencies

SSD-based

- Supports concurrent access from hundreds of thousands of cores

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.
S3-linked file systems: Objects stored in S3 appear on FSx file systems.

- s3://bucket/file1.txt
- s3://bucket/file2.txt
- s3://bucket/folder1/file3.txt
- s3://bucket/folder2/file4.txt

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.
S3-linked file systems: Objects stored S3 appear on FSx file systems

- `s3://bucket/file1.txt` maps to `file1.txt`
- `s3://bucket/file2.txt` maps to `file2.txt`
- `s3://bucket/folder1/file3.txt` maps to `folder1/file3.txt`
- `s3://bucket/folder2/file4.txt` maps to `folder2/file4.txt`
S3-linked file systems: Objects stored S3 appear on FSx file systems

- s3://bucket/file1.txt
- s3://bucket/file2.txt
- s3://bucket/folder1/file3.txt
- s3://bucket/folder2/file4.txt
- s3://bucket/folder2/file5.txt

- /file1.txt
- /file2.txt
- /folder1/file3.txt
- /folder2/file4.txt
- /folder2/file5.txt

Amazon S3

Amazon FSx for Lustre

Amazon EC2

Amazon EC2

Amazon EC2
S3-linked file systems: New files can be exported from FSx to S3

- Amazon S3
  - `s3://bucket/file1.txt`
  - `s3://bucket/file2.txt`
  - `s3://bucket/folder1/file3.txt`
  - `s3://bucket/folder2/file4.txt`
  - `s3://bucket/folder2/file5.txt`

- Amazon FSx for Lustre
  - `/file1.txt`
  - `/file2.txt`
  - `/folder1/file3.txt (edit)`
  - `/folder2/file4.txt`
  - `/folder2/file5.txt`
  - `/folder2/file6.txt (new)`

- Amazon EC2
- Amazon EC2
- Amazon EC2
S3-linked file systems: New files can be exported from FSx to S3

- s3://bucket/file1.txt
- s3://bucket/file2.txt
- s3://bucket/folder1/file3.txt
- s3://bucket/folder2/file4.txt
- s3://bucket/folder2/file5.txt
- s3://bucket/folder2/file6.txt

- /file1.txt
- /file2.txt
- /folder1/file3.txt (edit)
- /folder2/file4.txt
- /folder2/file5.txt
- /folder2/file6.txt (new)

Amazon S3

Amazon FSx for Lustre

Amazon EC2
S3-linked file systems: Spin up / spin down with compute resources

Amazon S3

s3://bucket/file1.txt
s3://bucket/file2.txt
s3://bucket/folder1/file3.txt
s3://bucket/folder2/file4.txt
s3://bucket/folder2/file5.txt
s3://bucket/folder2/file6.txt

Amazon FSx for Lustre

/file1.txt
/file2.txt
/folder1/file3.txt (edit)
/folder2/file4.txt
/folder2/file5.txt
/folder2/file6.txt (new)

Spin down resources between workloads

Amazon EC2

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark.
Key services and hardware that enable HPC on AWS

- **Compute**
  - Amazon EC2
  - Bare metal performance
    - AWS Graviton 2
    - EC2 Ultra Clusters
- **Networking**
  - Elastic Fabric Adapter (EFA)
  - MVAPICH2-X-AWS is designed for EFA
- **Storage**
  - Amazon FSx for Lustre
- **Workflow**
  - AWS ParallelCluster
- **AWS Batch**
AWS ParallelCluster

Set one configuration file with parameters and AWS ParallelCluster will provision an elastic HPC cluster on AWS for you.

- Login Node (place to submit jobs, edit files, etc.)
- Compute Nodes (Scalable set of compute resources (increases and decreases with workload in queue))
- Shared Filesystem (Filesystem mounted on login node and all compute nodes)
- Make Viz Easy on login node

+ Add your own changes in either a base image, post-install script, or interactively (e.g., add MVAPICH2-X-AWS, specific tuning, application setup)
Let’s try it!
Key services and hardware that enable HPC on AWS

Compute
- Amazon EC2
- Bare metal performance
  - AWS Graviton 2
  - EC2 Ultra Clusters

Networking
- Elastic Fabric Adapter (EFA)
- MVAPICH2-X-AWS is designed for EFA

Storage
- Amazon FSx for Lustre

Workflow
- AWS ParallelCluster
- AWS Batch
AWS Batch Overview

1. **Job Definition**
   Template that has common attributes (container image, IAM role, vCPU & memory requirements, …)

2. **Job**
   Each job must reference a job definition, but many parameters may be overridden when submitted

3. **Job Queue (JQ)**
   Queue determines priorities. Each JQ is connected to 1 or more CE

4. **Compute Environment (CE)**
   Resource Mix (defines On-demand vs. Spot and instance types. CE can be connected to more than one JQ

---

Jobs → Job Definition → Jobs

Queue → Queue → Compute Environment

Container → Container → Compute Environment

Store Results → Store Results
Key services and hardware that enable HPC on AWS

- **Compute**: Amazon EC2
- **Networking**: Elastic Fabric Adapter (EFA)
- **Storage**: Amazon FSx for Lustre
- **Workflow**: AWS ParallelCluster, AWS Batch
Our team has customers on Mars ...

We’re hiring LOTS of HPC Software Developers (and managers) in Seattle and Asti (in Italy) and will offer relocation packages to the right candidates as well as attractive comp and an incredibly fun work environment.

http://hpc.news/jobs

WE’RE HIRING ... to change the world(s).
Thank you!

mkoop@amazon.com