Achieving Cost-efficient, Data-intensive Computing in the Cloud

Michael Conley¹, Amin Vahdat²,¹, and George Porter¹
University of California, San Diego¹ and Google Inc.²

Configuring Cloud Deployments

- **Problem:** Picking a cost-efficient, large-scale deployment in the public cloud is challenging
  - Amazon EC2 has 53 VM types
  - Performance doesn’t match spec sheet
  - Performance decreases at scale

- **Consequences**
  1. Picking the wrong VM substantially increases cost - 100x in the worst case
  2. Sub-linear scaling further increases cost - Up to 40% in some cases

- **Key Insight:** Measuring VM performance at scale allows for accurate performance and cost prediction
  1. Restrict scope to I/O-bound workloads
  2. Build simple application model
  3. Measure storage and network at scale
  4. Compute cost-efficient configuration

Measuring Amazon Web Services

- Compute job cost under three I/O assumptions
  1. Modeled storage performance only (small-scale)
  2. Modeled network and storage (small-scale)
  3. Measured network performance (large-scale)

Benchmarking Storage and Network

- **DiskBench**

  - Input Disks → Distribute Partitions → Output Disks
  - $B_{read}$ → $B_{write}$ → $B_{storage}$

- **NetBench**

  - Generate Synthetic Data → Shuffle Data → Delete Data
  - $B_{network}$

Sort 100TB

- Annual 100TB GraySort benchmark competition
  - 186 i2.8xlarge with placement groups
  - Set Daytona world record – 4.35 TB/min

- New 100TB CloudSort benchmark
  - Public cloud with persistent storage
  - 330 r3.4xlarge with placement groups and EBS
  - Set Indy/Daytona world record - $451