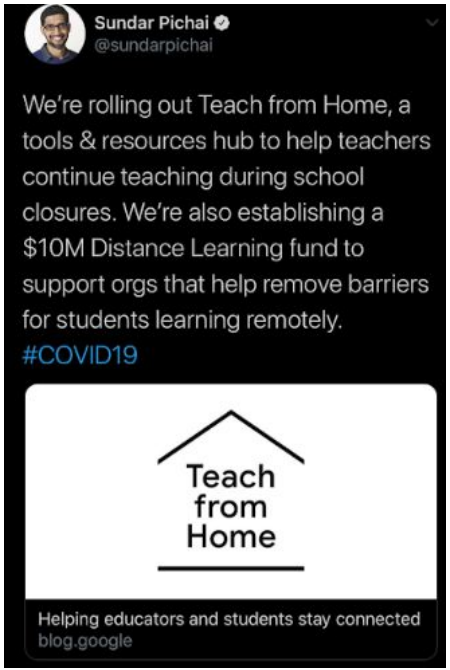
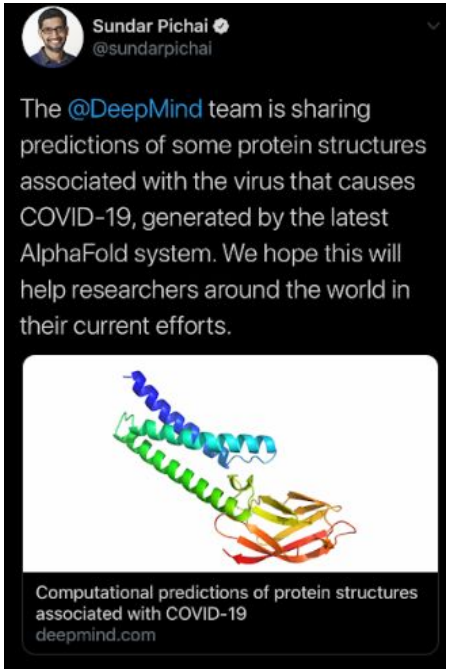
Google Cloud - ATLAS

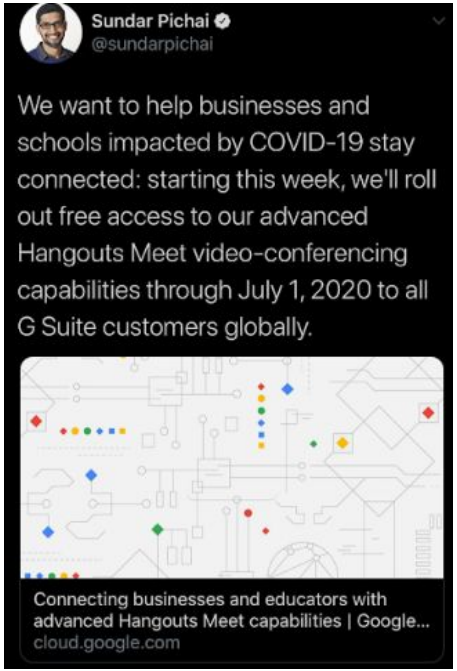
Presented by:
Karan Bhatia, SA Manager, karanbhatia@google.com
Usman Qureshi, CE, googleusman@google.com

Google Cloud

What have been up to the past week...







Google Cloud

Our products are how we innovate

1997 Search
Search ranking
Speech recognition

2005 Maps
Street View image
Popular times

2008 Android
Keyboard input
Assistant

2012 Play
Recommendations
Developer mode

2015 Photos
Photo search
Smart sharing

2004 Gmail
Smart Reply
Spam classification

2006 YouTube
Recommendations
Better thumbnails

2008 Chrome
Hack detection
Translation

2012 Drive
Quick Access
Data Loss Prevention

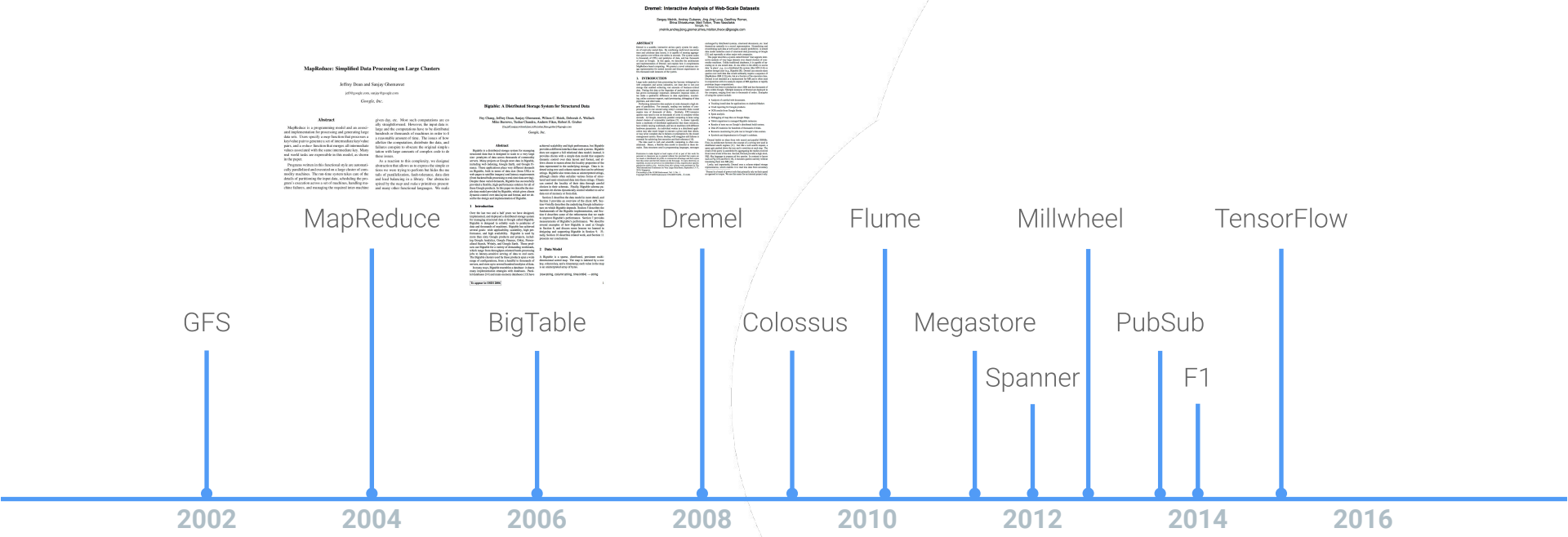
Google Cloud

9
Billion User Applications

100%
Machine Learning Powered

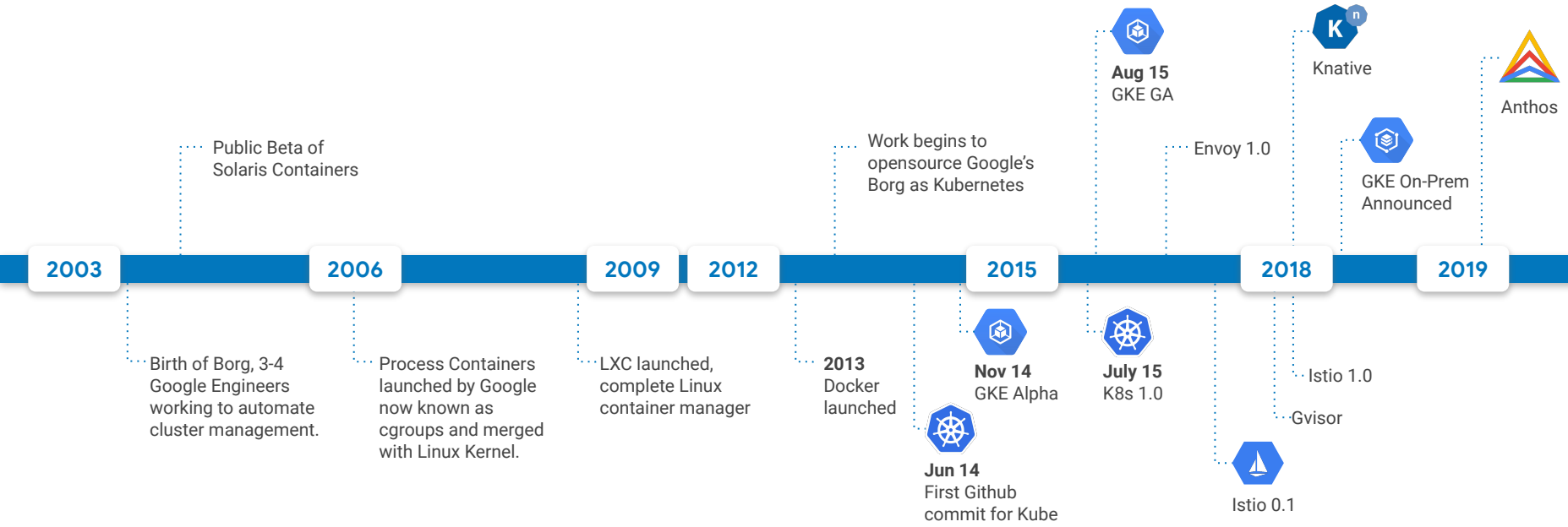

Google Cloud
$29 billion investment in the last 3 years
Over 1 billion unique IP addresses served daily
5,000 caching points across the globe
Google Compute Engine guarantees 99.95% uptime
Google Cloud

15+ years of data research at Google



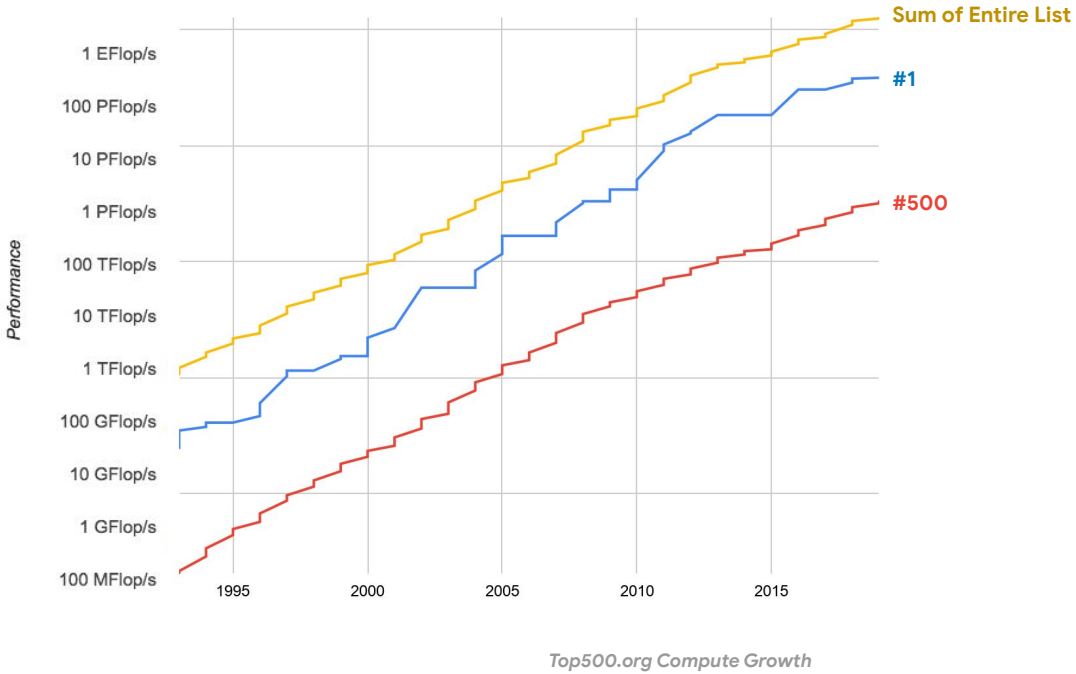
Google Cloud

15+ years experience in application orchestration

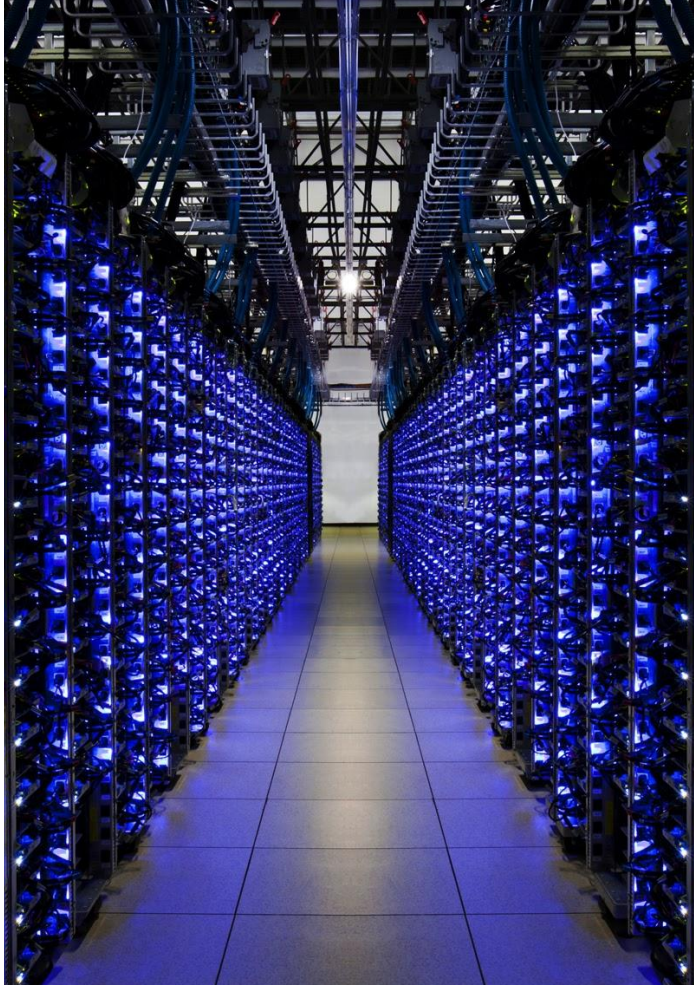


Today, on average, we launch 4 billion containers per week

Compute needs are growing



Top500.org Compute Growth



Data is now BIGGER than big

2008		2018
Los Alamos National Laboratory - RoadRunner		Oak Ridge National Laboratory - Summit
3PB Capacity - Panasas Storage	**83x**	250PB Capacity - IBM Spectrum Scale
55 GB/s throughput	**45x**	2.5 TB/s throughput

Democratize high performance computing and make it universally accessible and useful.

We continue to make investments in HPC

slurm workload manager

Auto-scaling and hybrid features built for Google Cloud

Mar 2019

DDN STORAGE

DDN releases Cloud Edition for Lustre on GCP, places #8 on IO500

Apr 2019

elastifile

High-performance, scalable file storage, acquired by Google

Aug 2019

rescale

Popular HPC platform providing a GUI to run HPC workloads on GCP

Nov 2019

Apr 2019

NVIDIA

The first major provider to offer T4 GPU globally

Apr 2019

intel

Computed Optimized VMs (C2) announced. 40% higher performance than N1 VMs

Aug 2019

AMD

EPYC VMs (N2D) announced. Up to 60% better platform memory bandwidth than existing instances

Nov 2019

Batch on GKE

Cloud-native workload manager for containerized HPC batch jobs

Google Cloud

Outcomes

Clemson broke the world record for running an HPC workload using the most compute cores on any commercial cloud.

2.138M vCPU | 133,573 instances | 200+ TB data processed

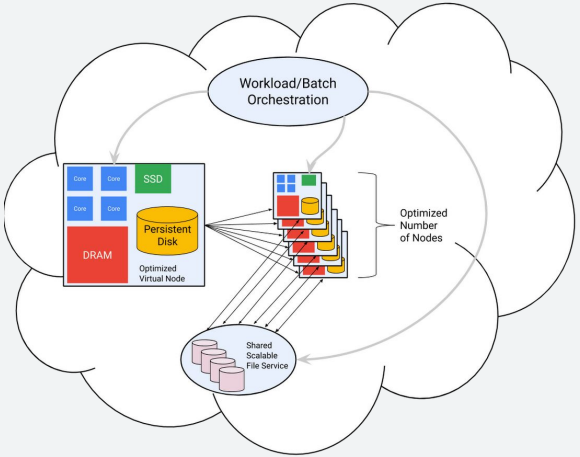
Average cost of **$0.008 USD per vCPU hour**

Published **"On-Demand Disaster Management using High Performance Computing in the Commercial Cloud,"** in The Journal of Supercomputing

A corpus of traffic data for public use to be added to the GCP Marketplace via the Public Dataset Program

[Google Cloud Blog Post]

Google Cloud



Computing

Google Cloud

Google Compute Engine

Custom Machine Types

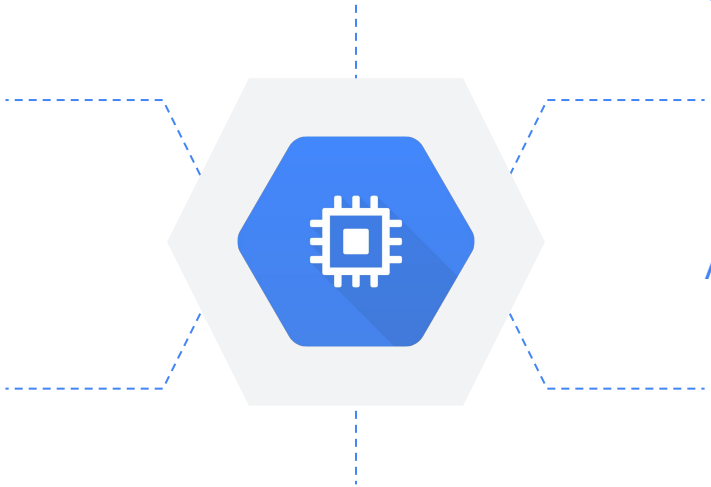
- Choose your Core:RAM ratio
- Recommendation engine for optimum machine size

Compute power

- Up to 416 vCPUs
- Up to 12TB RAM
- Up to 3TB Local SSD
- Up to 8 GPUs
- Up to 7TB Intel Optane

Compute Optimized VMs

- Fast Clocks - 3.8 GHz
- Up to 60 vCPU, 240GB RAM
- vNUMA exposure
- Overhead reduction



Networking

- Globally routed subnets default
- 2Gbps per vCPU, up to 32Gbps
- 15,000 VMs per Virtual Private Cloud (VPC)

Accelerators

- Up to 8 NVIDIA GPUs
- NVIDIA K80, P100, P4, V100, T4
- NVLink and PCIe x16 for maximum performance
- Google Cloud TPUs for ML

Preemptible VMs

- Up to 80% discount
- Custom Machine Types, GPUs, Local SSDs

NVIDIA GPUs

Attach up to eight GPUs per instance

NVLink and PCIe x16 to achieve maximum performance

Per-second billing, Sustained Use Discounts
Preemptible support (~70% off)

NVIDIA Tesla K80
NVIDIA Tesla P100
NVIDIA Tesla P4
NVIDIA Tesla V100
NVIDIA Tesla T4

Google Cloud

“The Jungle Book” wins Best VFX Oscar in 2017.

Rendered on Google Cloud Platform.

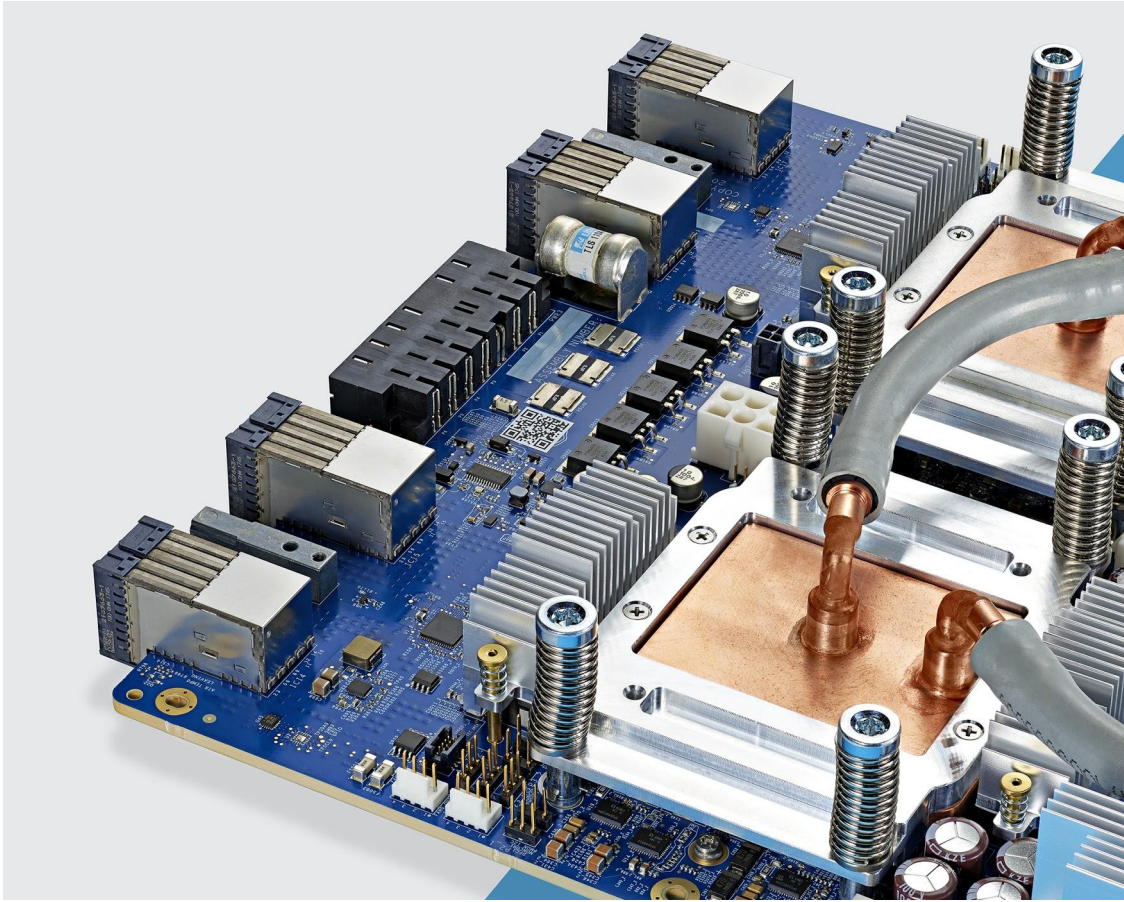
5.5 million core hours
41 years of compute time over two months
1.5 million tasks processed

(MPC film, 2017)

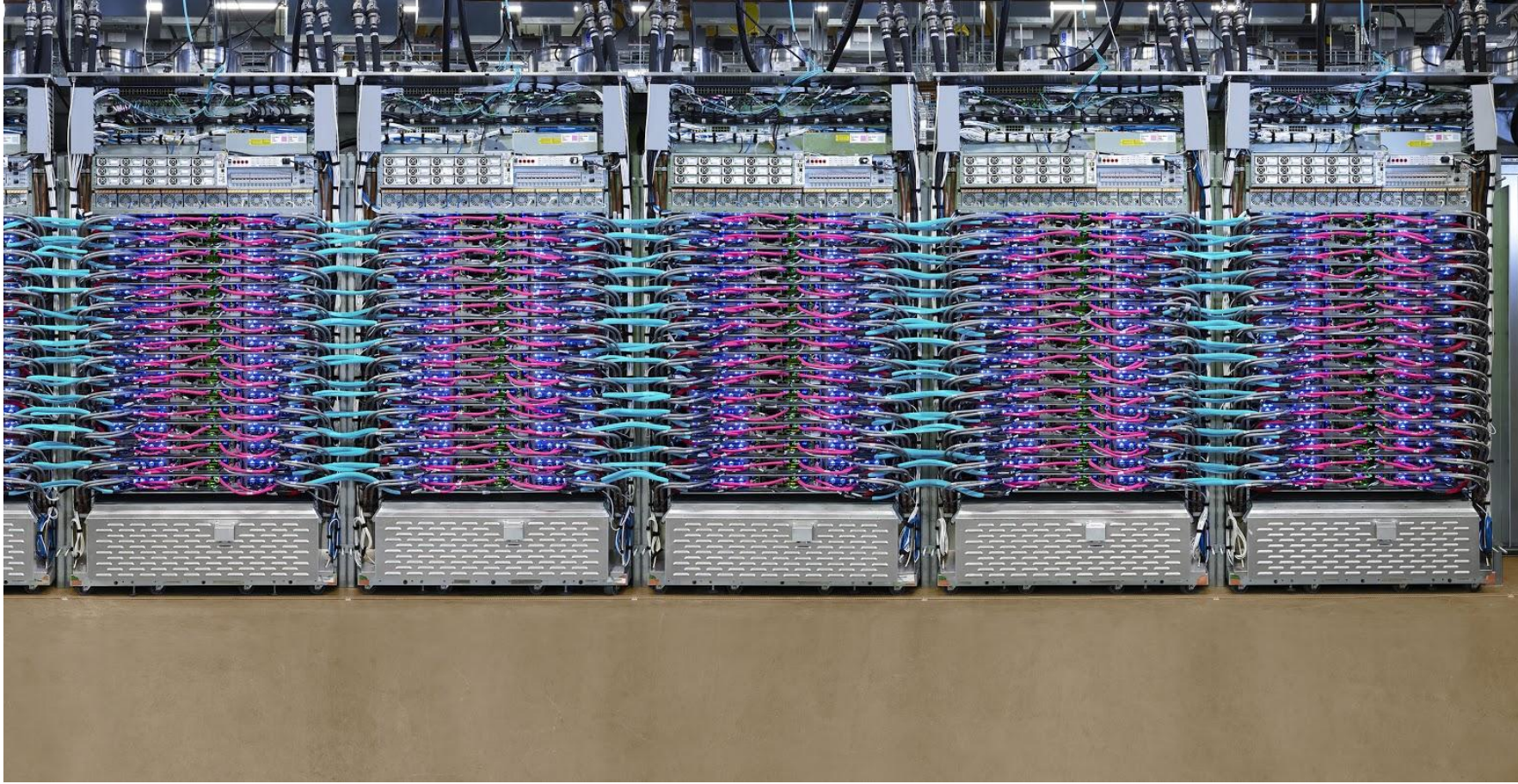
Google Cloud

Cloud TPU v3

420 teraflops
128 GB of high-bandwidth memory
Training and inference
Generally Available (GA)



Google Cloud



Google Cloud

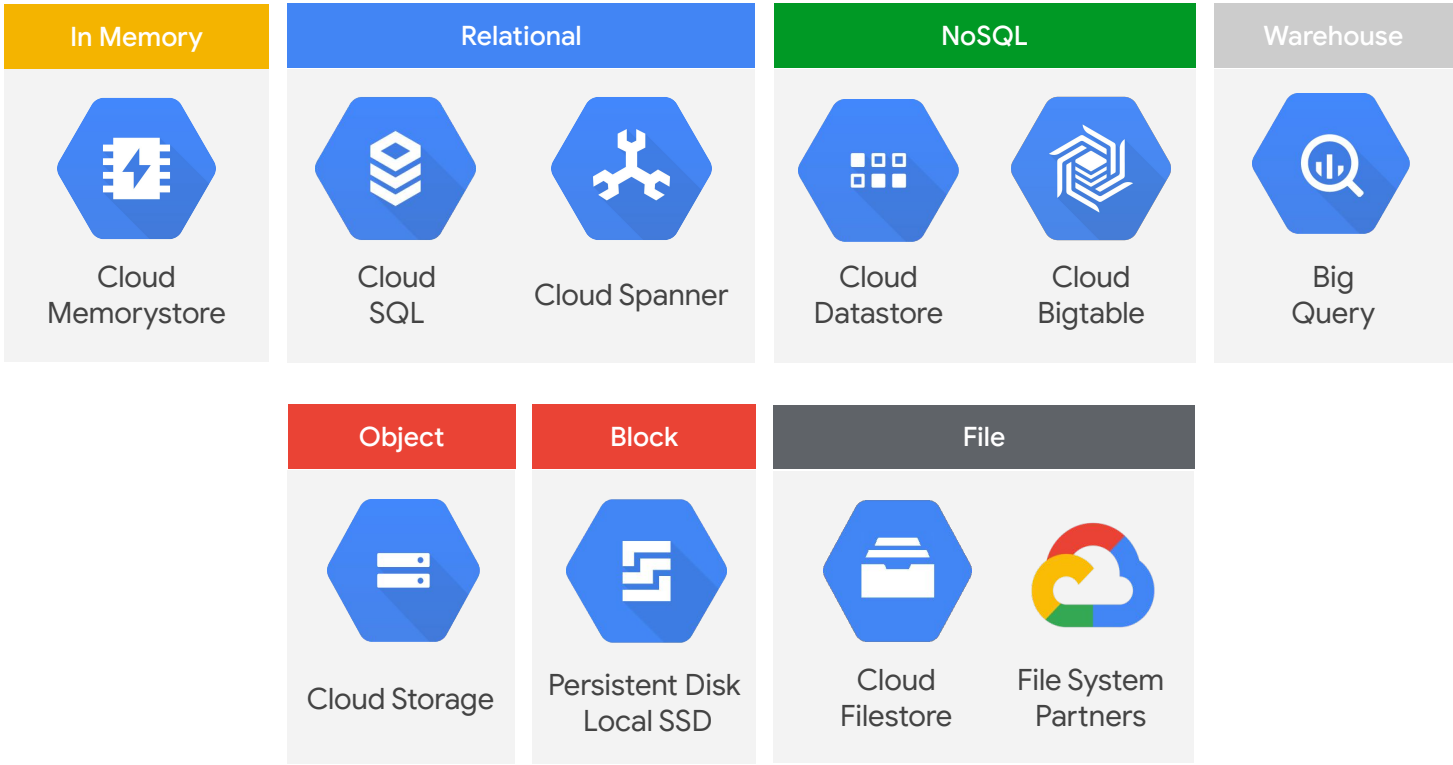
Cloud TPU v3 Pod: >100 PFLOPS

(8x faster than v2 Pod)

Storage

Google Cloud

Where do I store my data in Google Cloud?



Google Cloud

HPC Storage on GCP

Google Cloud Storage
Exabyte-scale, feature-rich object storage
Automatically scaling throughput

Persistent Disk
SSD/HDD Persistent Disk
High-performance, replicated block storage

Local Storage
Local SSD (NVMe) for scratch and fast access
Physically attached to node via PCI

Cloud Filestore
Highly available, durable, POSIX-compliant shared storage across tens of thousands of nodes

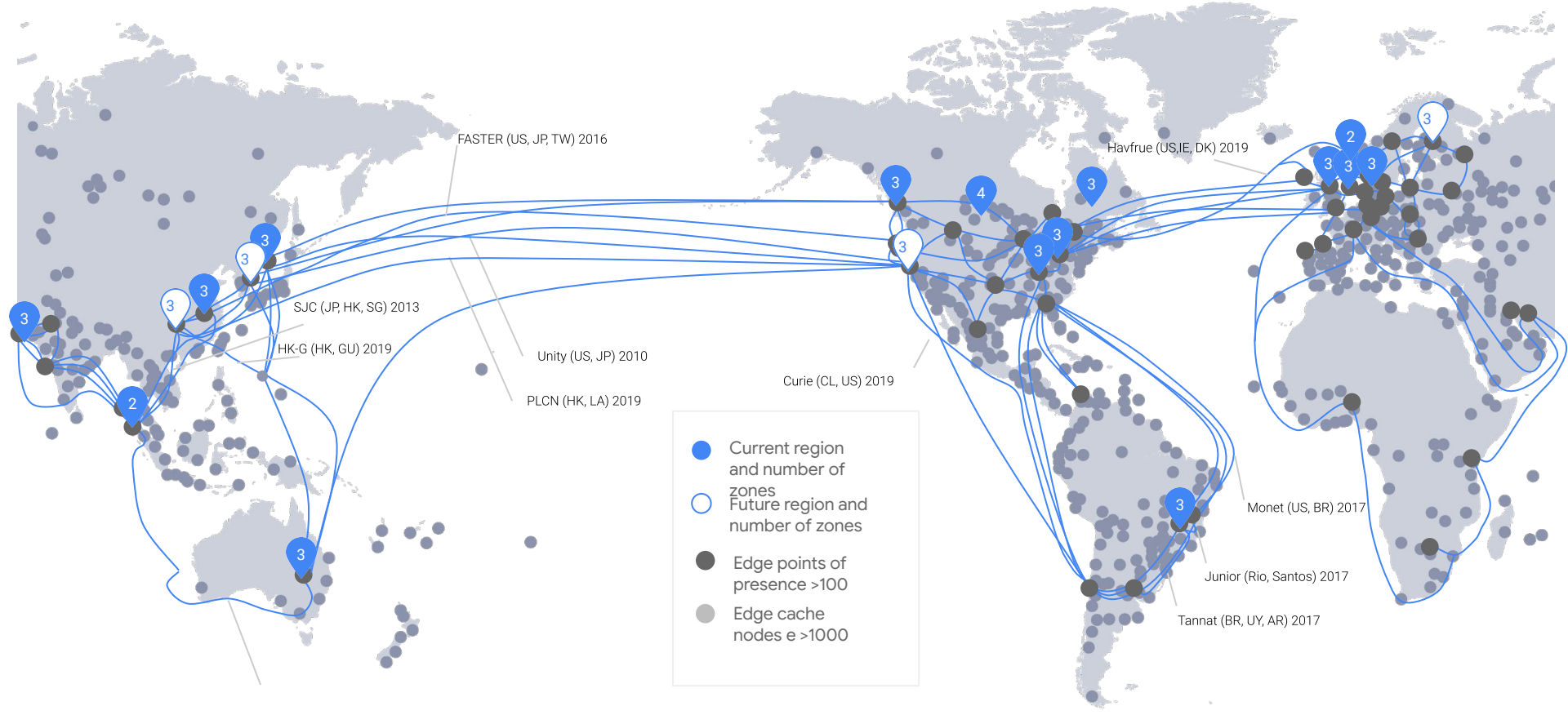
Partner, hybrid, and open-source
Storage solutions for NetApp, Elastifile, DDN, Lustre, and more
Move petabytes to GCS with the Data Transfer Appliance

Google Cloud

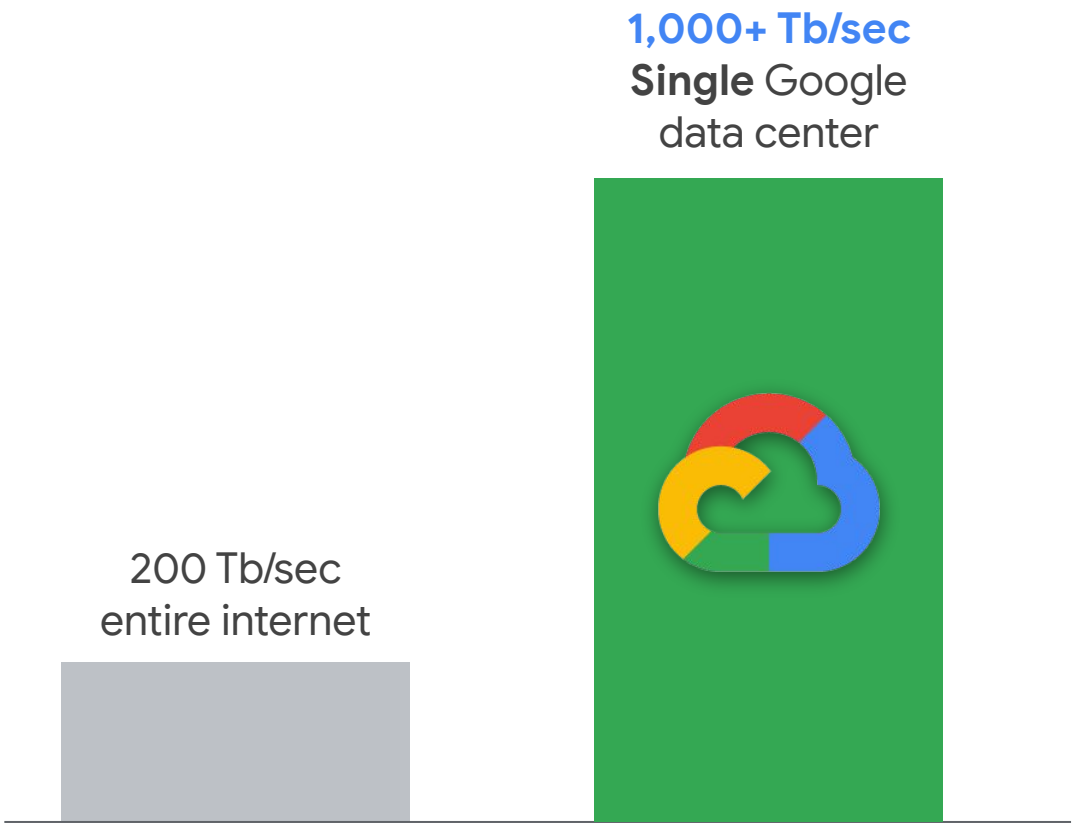
Networking

Google Cloud

Global network infrastructure



Bisectional bandwidth



Google Cloud

https://cloudplatform.googleblog.com/2015/06/A-Look-Inside-Googles-Data-Center-Networks.html

Google Networking

Performance

- 15,000 VMs per VPC
- Predictable, low latency (~20 - 40 µs)
- Scalable bandwidth
  - 2 Gbps per vCPU
  - Up to **32Gbps** per VM
- Open-sourcing and tailoring latency-sensitive tools to GCP (gRPC)

Global Network

- Thousands of POP around the world
- Google Backbone between datacenters
- Multiple interconnect options to on-prem

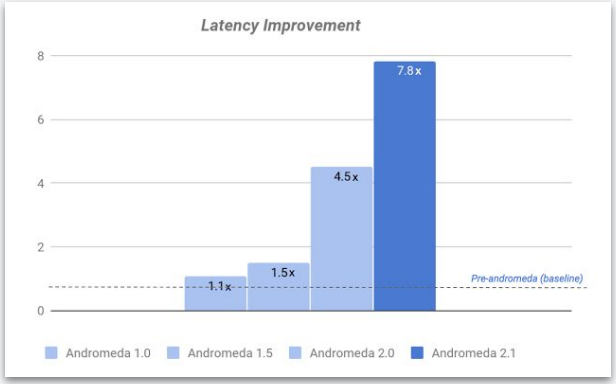
Google Cloud

Network

Network/
Bandwidth

Efficient SDN network

- Clos topology, a collection of smaller custom switches arranged to provide the properties of a much larger logical switch.
- Centralized software management stack.
- Relying more on custom protocols tailored to the high performance data center.



Batch on GKE *Preview*

Batch Job Management Platform built for Kubernetes; available on GKE

- Extends K8S APIs to add support for Batch Jobs and Job Queues.
- New Autoscaler and Scheduler optimized for batch workloads

Google Cloud

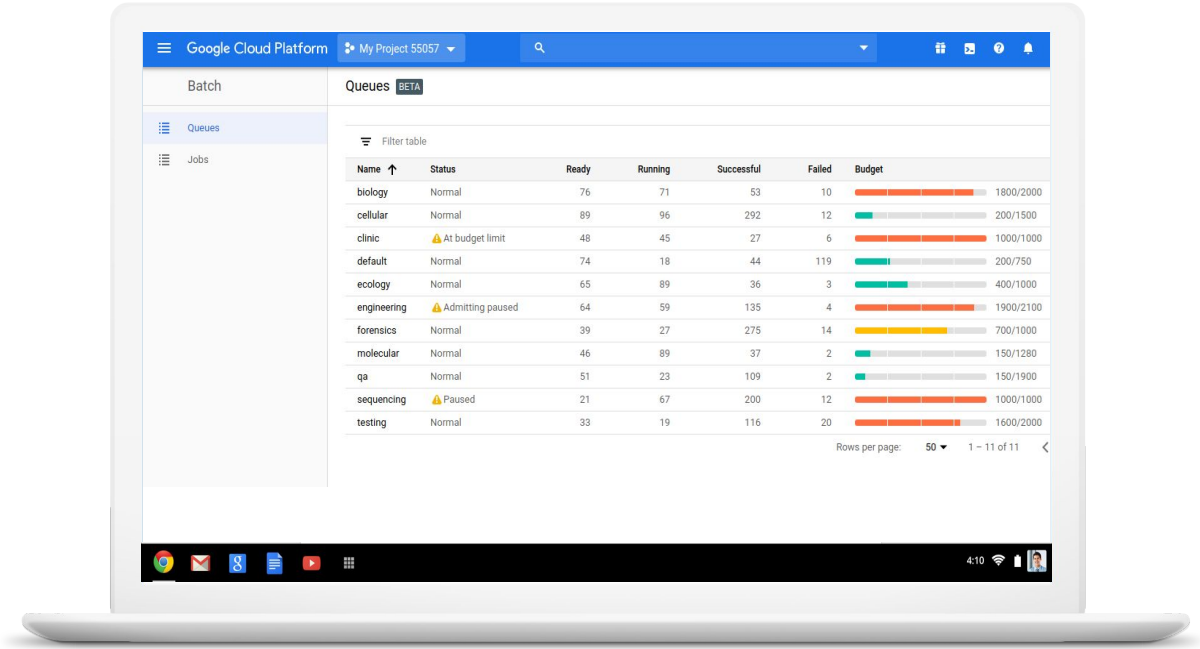
Features

Batch Constructs

1. Queues and Queue priorities
2. Job dependencies and constraints
3. Array Jobs

Cost Management

4. Budget Controls
5. Autoscaling
6. Rightsizing



Google Cloud

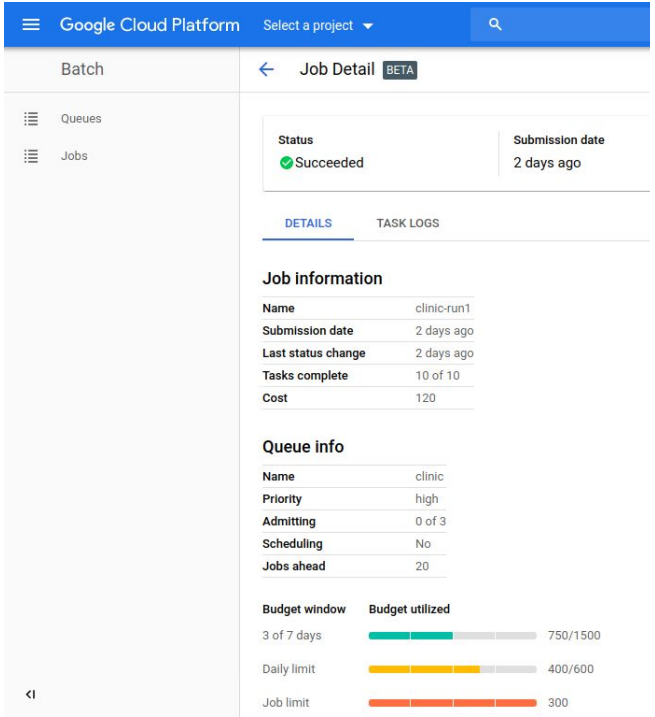
Features (Contd)

Ease of use

1. User account management
2. Data copy to/from cluster
3. Ksub command line tool for easy job submissions and management

Platform Integration

4. GPU support
5. Stackdriver Monitoring
6. Persistent job logs



Google Cloud

Thank you

Google Cloud