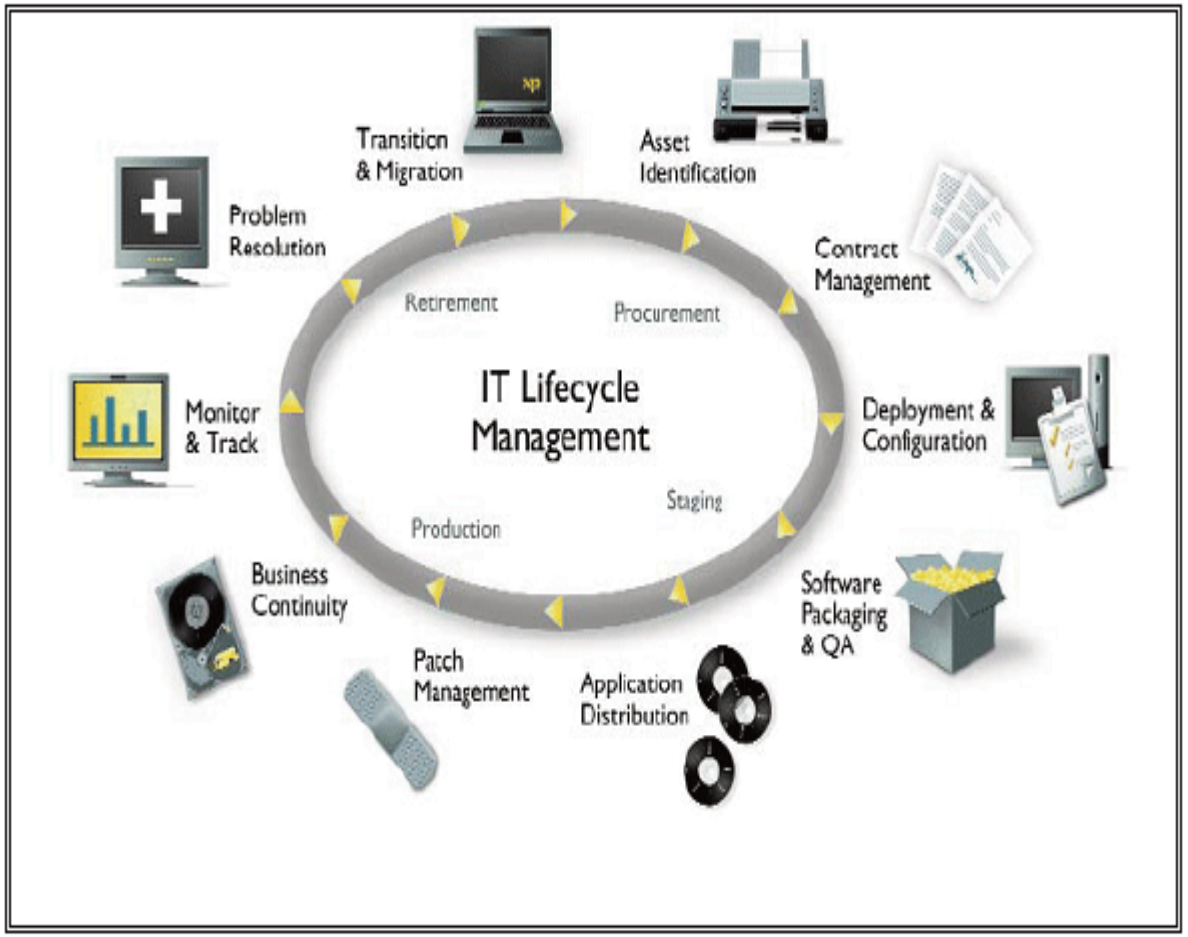
SNIA DSI '16 DATA STORAGE INNOVATION CONFERENCE
Innovation in Storage Products, Services, and Solutions
June 13-15, 2016 | Marriott San Mateo | San Mateo, CA

Cloud Bursting

Jim Thompson
Avere Systems

Why Use Someone Else's Cloud?



Significantly reduce infrastructure management costs both in money and time

Maintain operational flexibility during scale-out jobs…let the provider deal with scale challenges

Advantages of Using the Cloud for Compute

Run thousands of cores against thousands of (actuarial, financial, genomic, scientific, rendering) jobs

Run these cores on-demand

Enable the cores and jobs with a few lines of code or a script, or use your existing scheduling or dispatch infrastructure

When finished, execute a few more lines of code or script and tear it all down (if you like)

Relentless ability to repeat this

What type of Processing is Destined for Cloud Computing

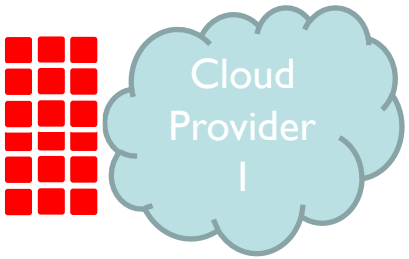
High Core Count Applications:

Computationally intensive: Monte Carlo simulations, Batch jobs, Rendering, Transcoding

Cache-able Read-heavy workloads: Reading in large amounts data that is read over and over again

Batch or burst usage model: Running simulations overnight

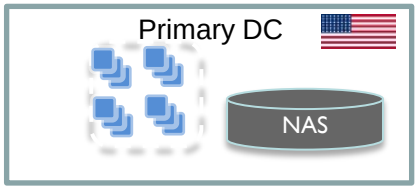
One Practical Approach to Leveraging Public Cloud



Locate work in cloud compute:

- Rendering
- Transcoding
- Simulation work

Use Cloud Services where applicable



Use local DC in two ways:

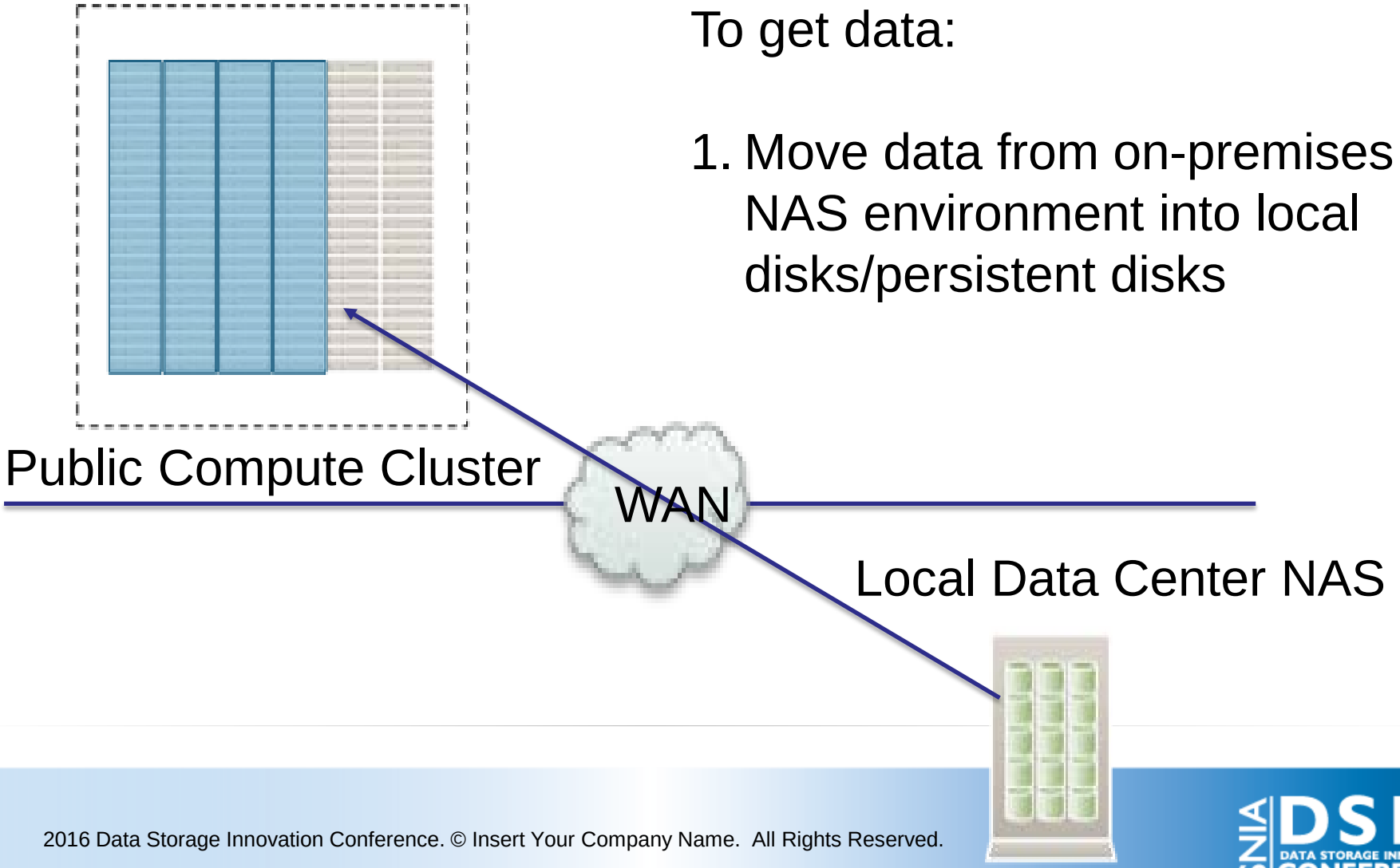
1. Until current assets are retired
2. Ongoing R&D / Best-in-Breed technology facility…tech that is ahead of the public cloud lifecycle

Implementation Considerations

Latency	Between Compute and Data
Locality of model data	Data sets are large and costly to move
Security / Regulatory	Data may need to be at rest in a specific geography Data encryption Restricted access to data
Cost	20,000 cpu cloud cores, 3 hours a day, annually ~$1.5M

Fundamental Question: How does cloud compute access data?

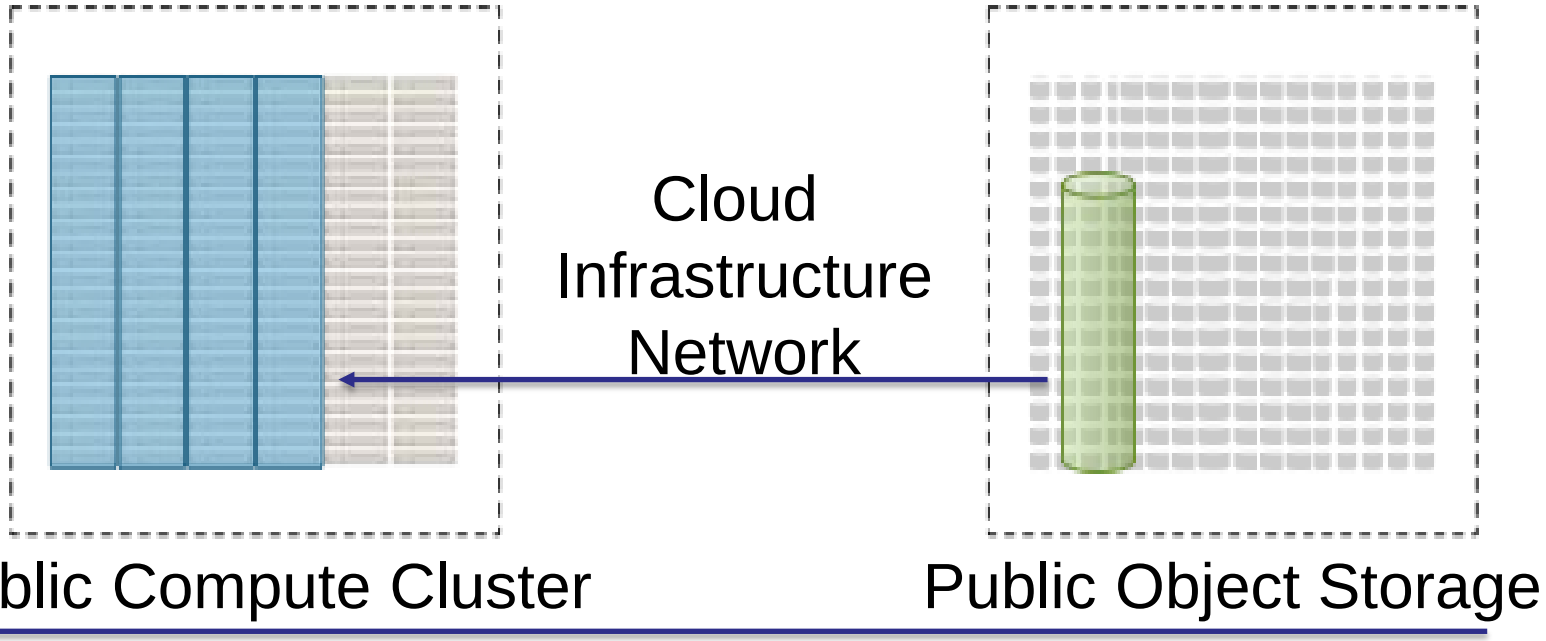
Cloud Compute Data Access Option 1



To get data:

1. Move data from on-premises NAS environment into local disks/persistent disks

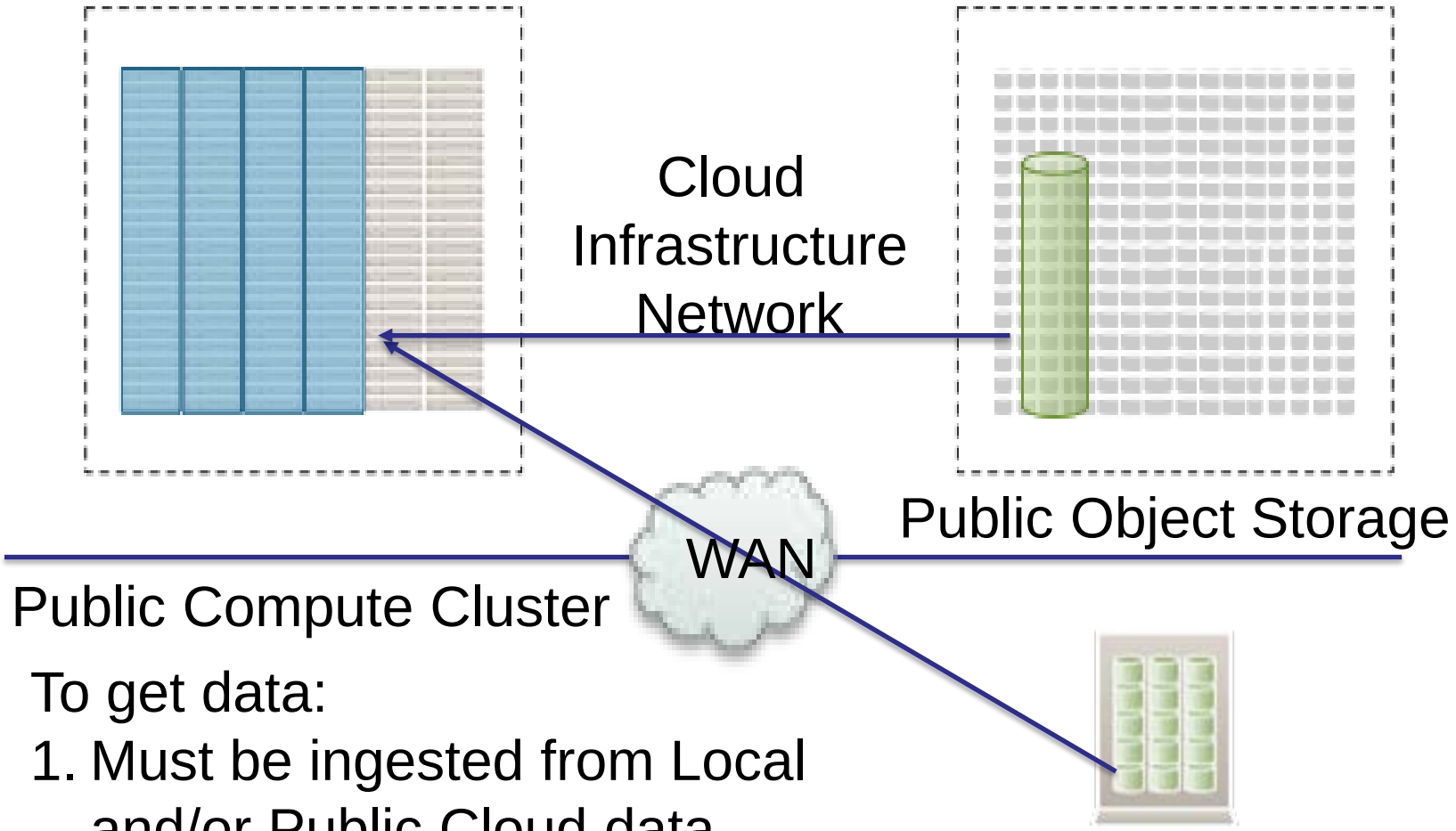
Cloud Compute Data Access Option 2



To get data:

1. Ingest Data from Cloud Provider Object Storage

Cloud Compute Data Access Option 3



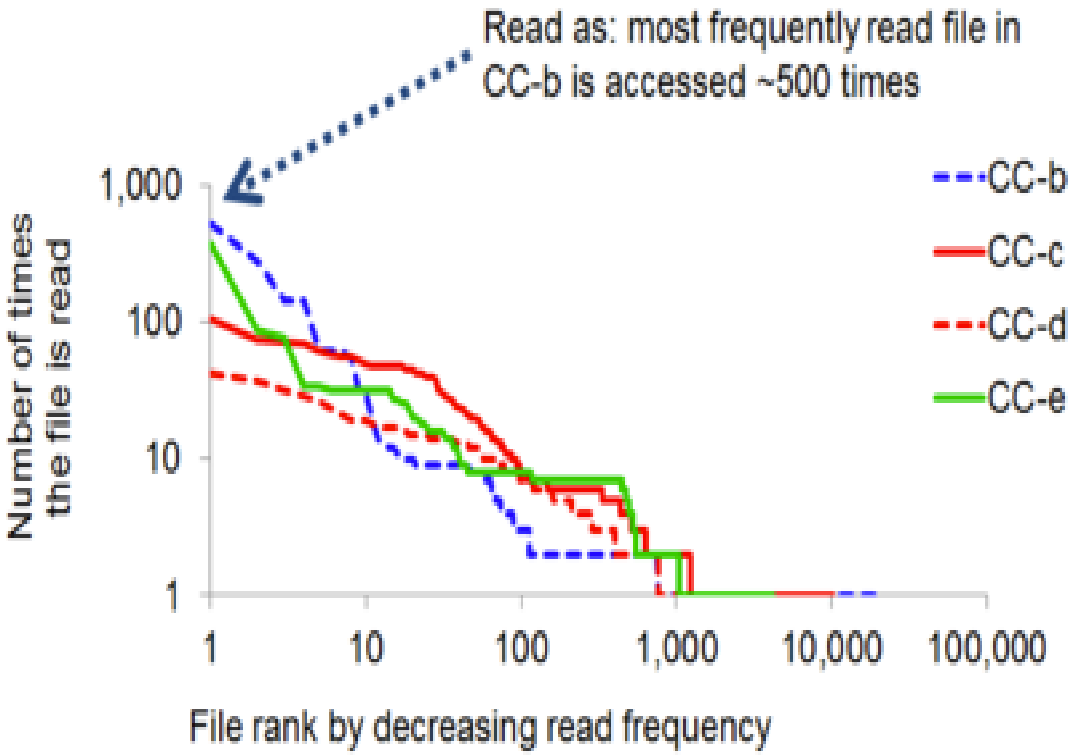
To get data:

1. Must be ingested from Local and/or Public Cloud data sources

Additional Challenges

- Data likely resides in multiple locations across WAN or Cloud Infrastructure connections
  - Increased Latencies for loading will result in longer ramp times for jobs
  - More susceptible to network issues

Typical File Access in Hadoop Cluster



Caching files will work for certain types of jobs

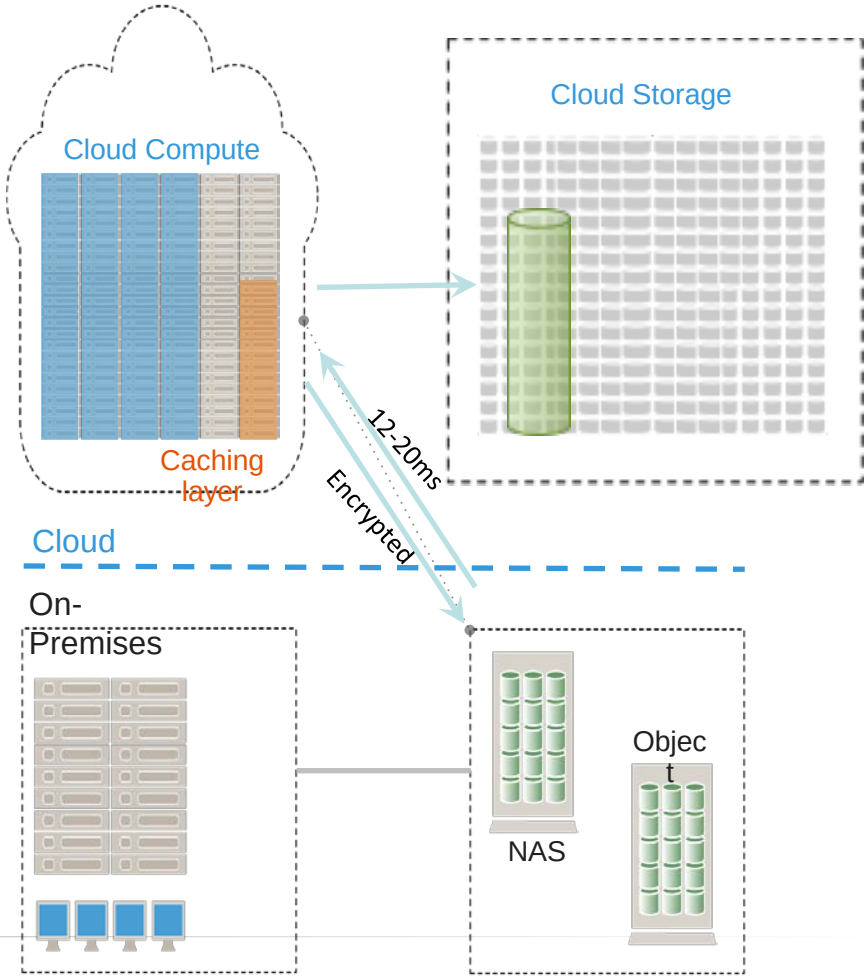
Where typical file is accessed by multiple clients

source: http://blog.cloudera.com/blog/2012/09/what-do-real-life-hadoop-workloads-look-like/

Need for a caching layer

Having caching layer to hold the repetitively requested data will allow each compute node to have instant access to the data, otherwise the compute nodes will need to traverse the WAN link each and every time the file(s) is accessed.

Data can be written back to on premise storage or into cloud storage.

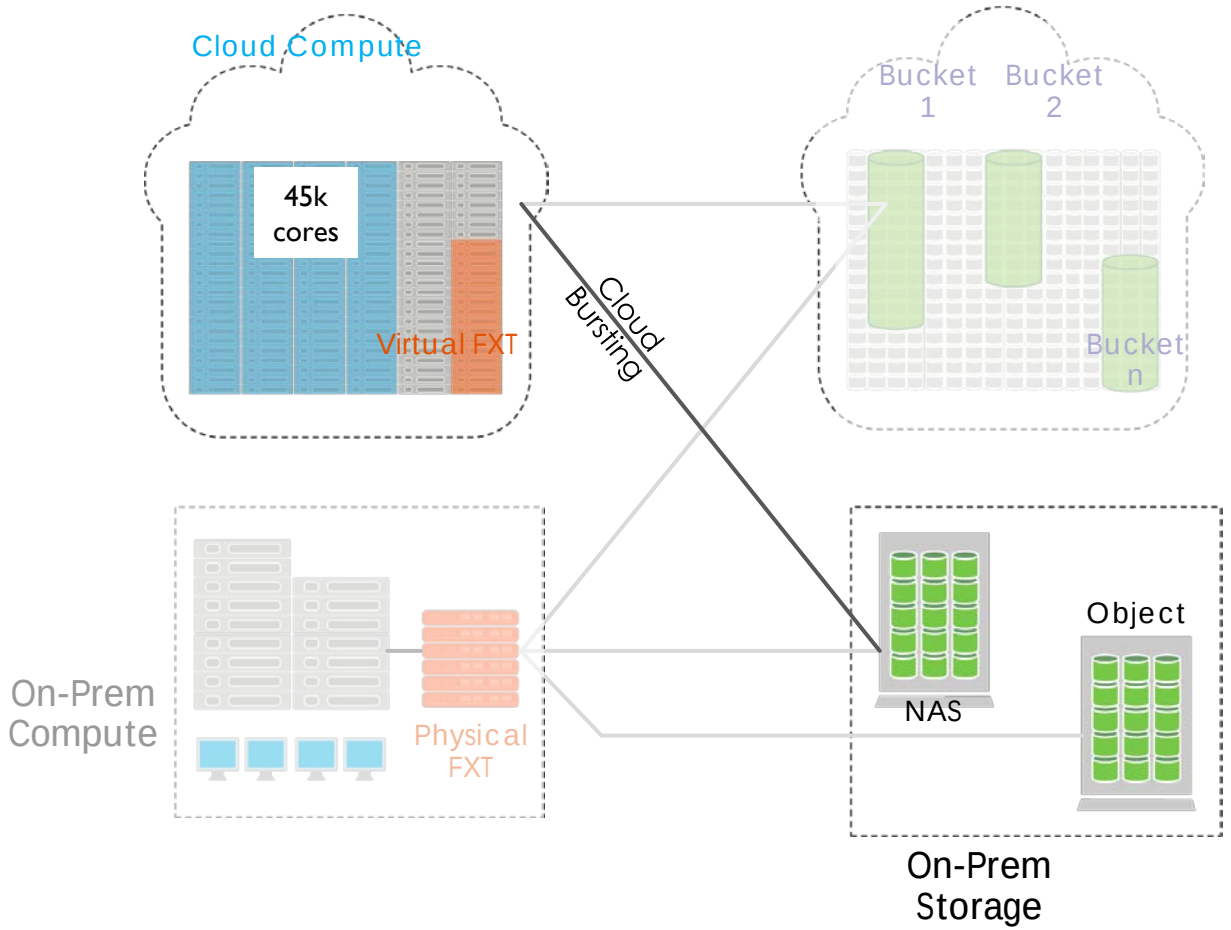


Use Case for Hybrid Cloud

Design goal: Run “wonderfully” parallel workloads on cloud compute, securely, economically, and at scale.

Requirements: data stays on-premises for daily access. Cloud compute needs to scale to double, triple and quadruple to what you have for on-premises compute to drastically reduce computational times. Applications need to be used “as they are”.

Use Case: Market Risk Analysis with Cloud Bursting



Customer Challenges

- Simulation complexity increasing
- Run 10-100x more simulations
- Finish models in less time
- Reduce $/simulation

Caching Benefits

- provides scalable file system for cloud
- Scale to more than 45k compute cores
- Auto-caching of data from on-prem storage
- Cloud economics: zero footprint, easy to turn on/off, pay only for what you use

Actual Customer Results

- 45k cores used in Major Cloud Provider

14

Questions?