Joint Software & IT-CAST Forum September 2020

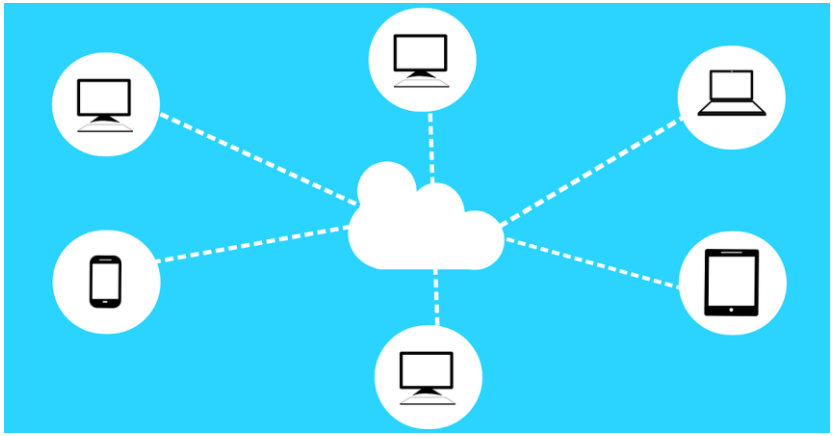
Developing CERs to Estimate Commercial IaaS Costs for Federal IT Systems

Cara Cuiule, PRICE Systems, LLC
Amanda Ferraro, PRICE Systems, LLC

Estimate with Confidence™

Overview

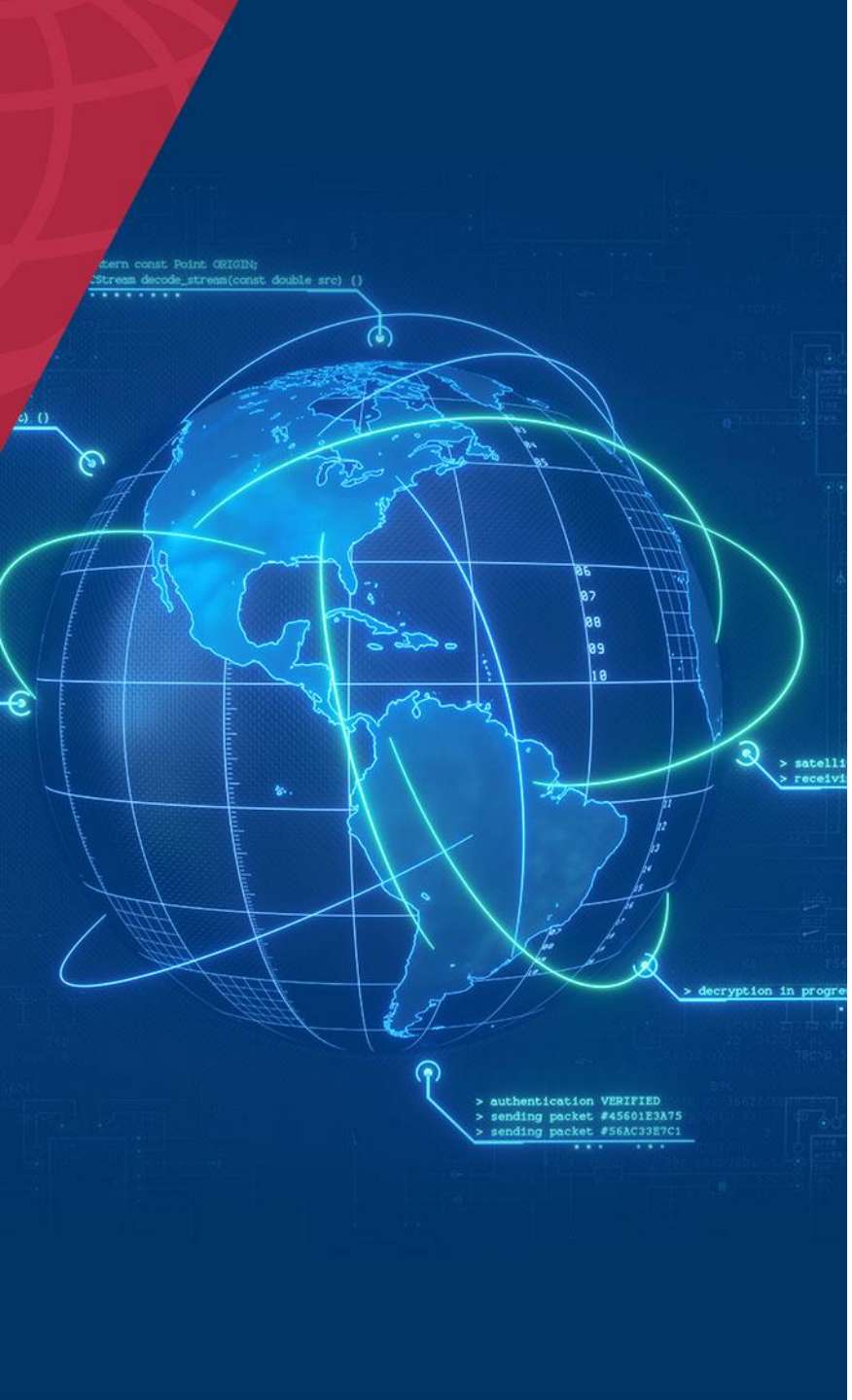
- Introduction
  - Purpose
  - Background
  - Research Focus
- Scope and Implementation
- Experimental Approach
  - Part I: Data
  - Part II: Virtual Machine CERs
  - Part III: Additional Storage Models
- Summary



Past Work

- Summer 2019:
  - Developed models for virtual machines and storage in the cloud

- Goals of update:
  - Collect additional data
  - Refresh models
  - Improve validation process



Introduction to Estimating Cloud Infrastructure Costs

Background

- Federal agencies are adapting their IT system budgets and funding strategies to accommodate commercial cloud computing (infrastructure as a service – IaaS)

- Defensible life-cycle cost estimates supporting IT budgets need to account for multiple types of Cloud services

- Multiple federal agencies have developed ad-hoc tools to estimate all or part of the cloud-related costs for specific providers

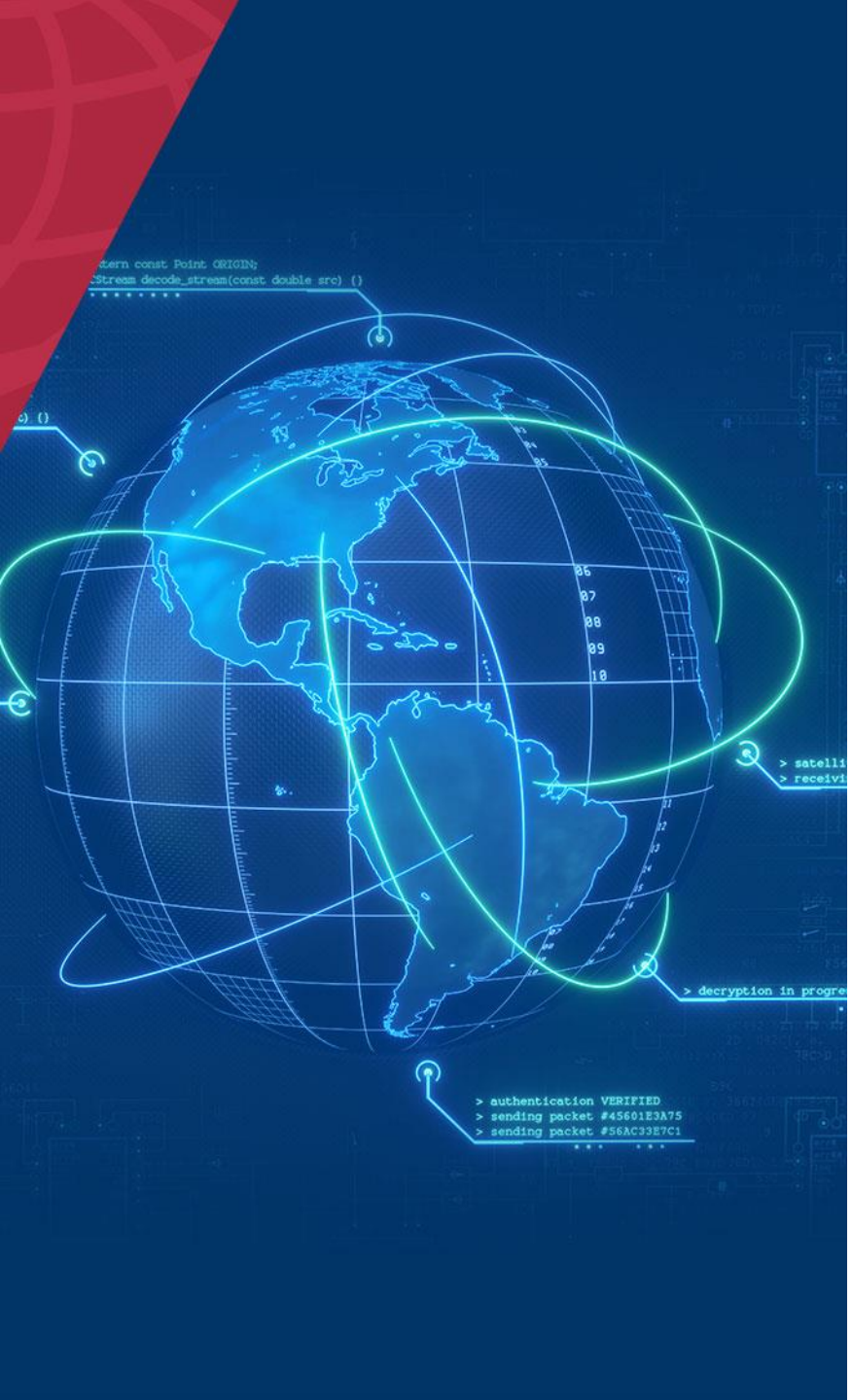
Issues

- Pricing structures and terminology for services are inconsistent between providers

- Federal Government cloud budgeting challenges:
  - Process is lengthy, complex
  - Initiated long before a cloud service provider is finalized
  - Difficult to predict budget requirements 1 to 3 years in advance of need

Solution

- This research briefing presents an approach to estimate verifiable IT Cloud costs using predictive analytic methods
  - Vendor agnostic; multiple vendors and instance options
  - Standardize pricing structures for various cloud services
  - Validated models using training/test subsets



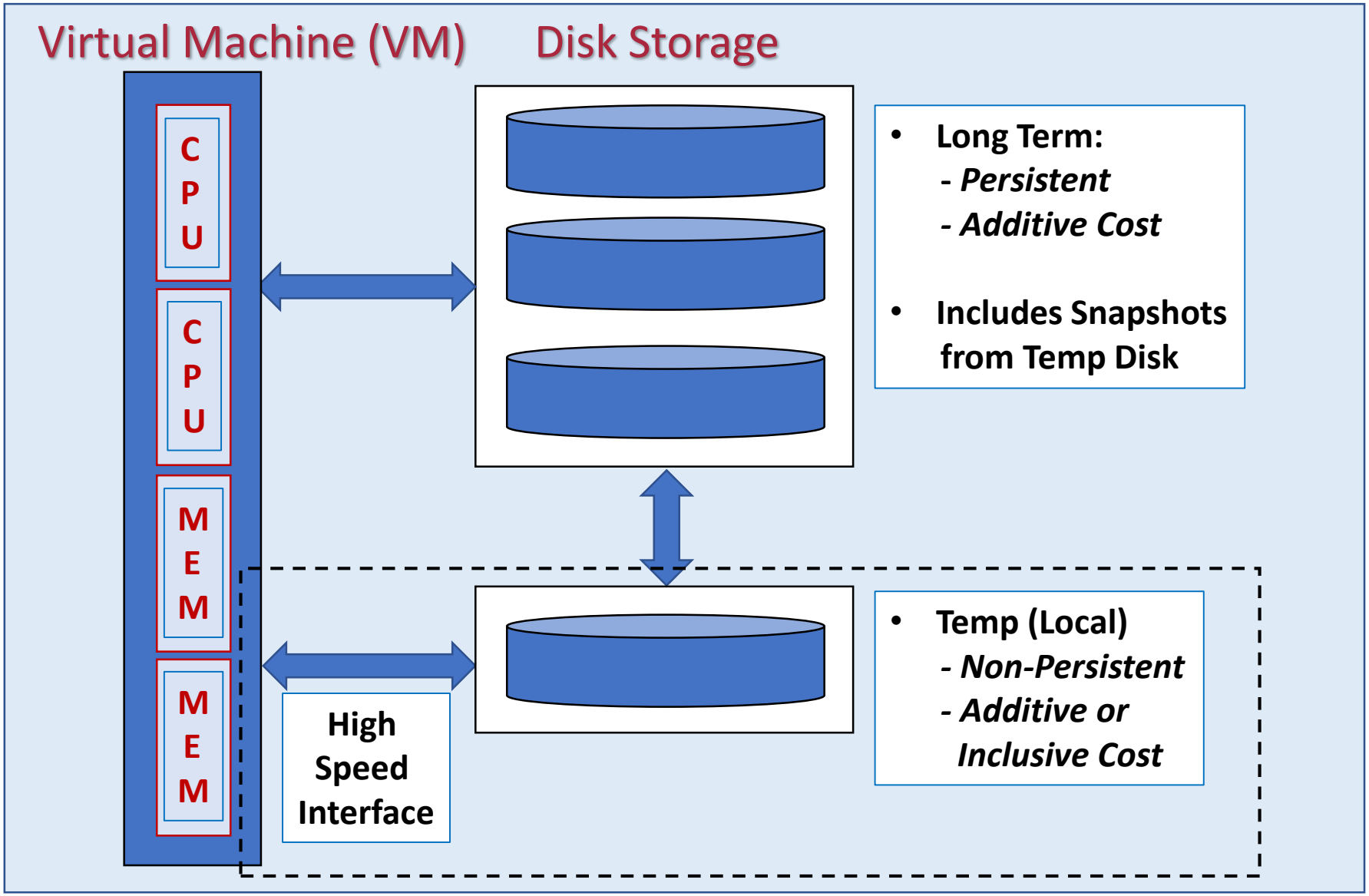
Scope and Implementation

Scope

- Data intended for federal business IT systems supporting finance, HR, medical, logistics, etc.
  - Open source cloud data only (no DISA), excludes private cloud instances
  - Virtual Machine (VM) and Storage pricing
  - 27,000+ world-wide datapoints extracted and normalized from 6 global service providers

- Part of an ongoing research effort on cloud service transition and operations costs

Focus on Infrastructure: Virtual Machines

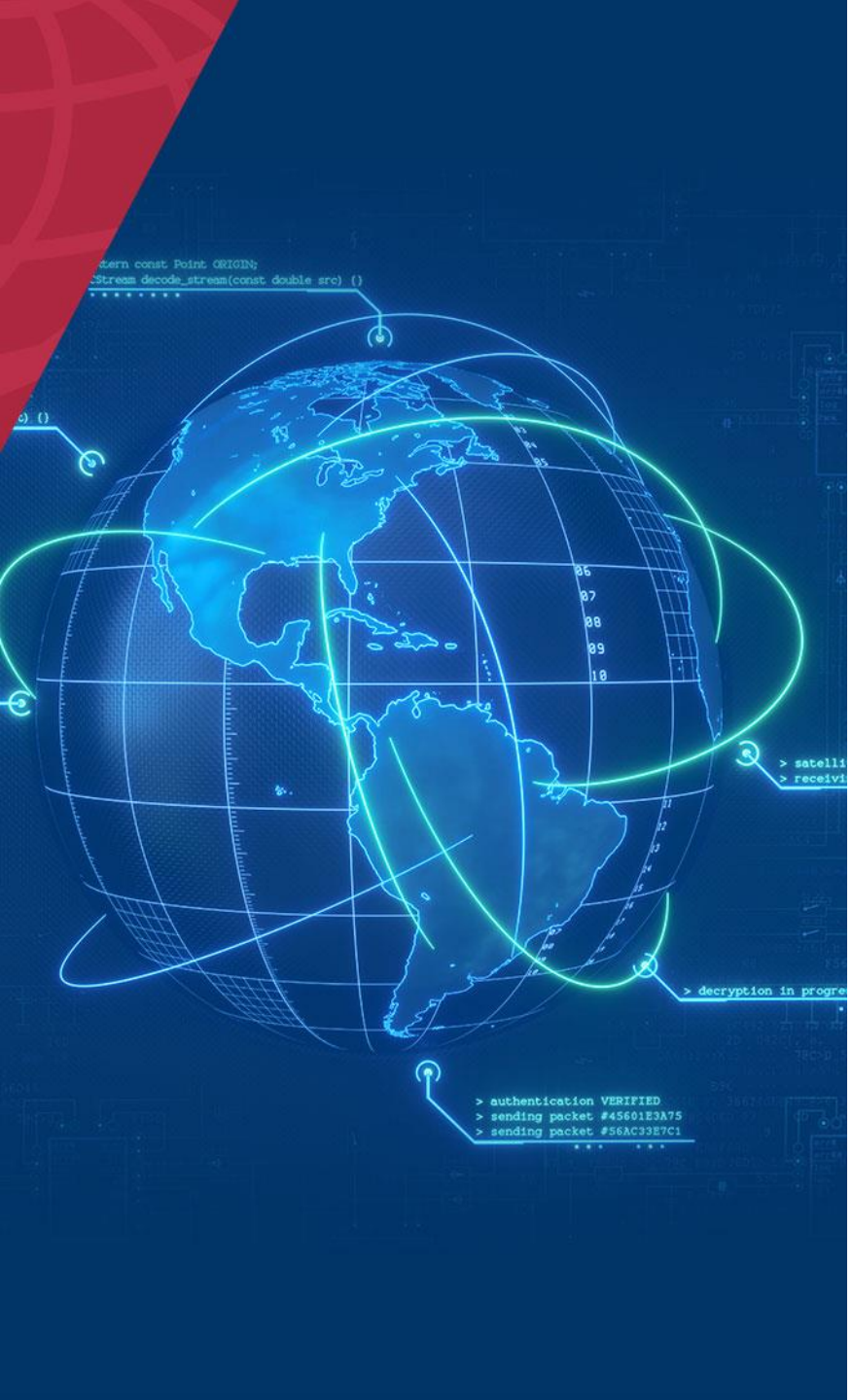


Focus on Infrastructure: Storage Types

- Object
  - Unstructured data (videos, photos, audio, etc.)
- File
  - Stores single pieces of data in folders for organization
- Block
  - Splits files into “blocks” of data and then stores these as separate pieces of data
  - Includes disks associated with VM instances (Local Disk)

Pricing Structures

Model	Infrastructure Components	Explanation
Basic Infrastructure Models		
Format 1	Number of CPUs + Memory (GBs)	Basic VM without Storage
Format 2	Added Storage	Basic Disk (Block, Object, File)
Combined VM and Storage Instances		
Format 3	CPUs + Mem + Temp Block Disk	VM + Included Storage; One Rate
Format 4	CPUs + Mem + Separate Object Disk	VM + Added Storage; Two Rates
Format 5	CPUs + Mem + Temp Block Disk + Snapshot Storage File	VM + Included Storage + Separate Snapshot; Two Rates



Experimental Approach

Part 1: Data

Sources

Source for commercial cloud cost and technical data:

- Open-source, non-proprietary data
- Individual Cloud Provider web sites and calculators

- 6 vendors: Amazon, Azure, Google, IBM, Oracle, Alibaba
  - *9,000+ USA datapoints*
  - *900 Gov. datapoints (VMs - Azure + Amazon, IBM included in Storage)*

- Additional sites
  - *Banzai Cloud*
  - *EC2Instances.info (Amazon Only)*
  - *Softlayer.com (IBM Cloud website)*

Categories - Virtual Machine Data

- Name
- Type
- CPUs
- RAM (Memory) (GB)
- GPUs
- GPU Type
- Local Disk (GB)
- Network Performance
- Provider
- Hourly Prices
  - On Demand Price
  - 1 Year Reserved Price
  - 3 Year Reserved Price
- Continent
- Region
- Location (City)
- Date
- Natural Log (ln) Transforms of Numerical Data

Name	Type	Memory	CPUs	GPUs	Local Disk Storage	Network Performance	Provider	On Demand Price
m5d.12xlarge	Compute	206.158081054...	48	0	1932.73205566406	10	Amazon	2.71199989318848
m5n.16xlarge	Compute	274.87744140625	64	0	0	75	Amazon	3.80800008773804
t3a.xlarge	Compute	17.1798400878...	4	0	0	35	Amazon	0.150399997830391
m5dn.xlarge	Compute	17.1798400878...	4	0	161.061004638672	25	Amazon	0.272000014781952
c5.18xlarge	Compute	154.618560791...	72	0	0	25	Amazon	3.05999994277954
m5ad.4xlarge	Compute	68.7193603515...	16	0	644.244018554688	10	Amazon	0.824000000953674

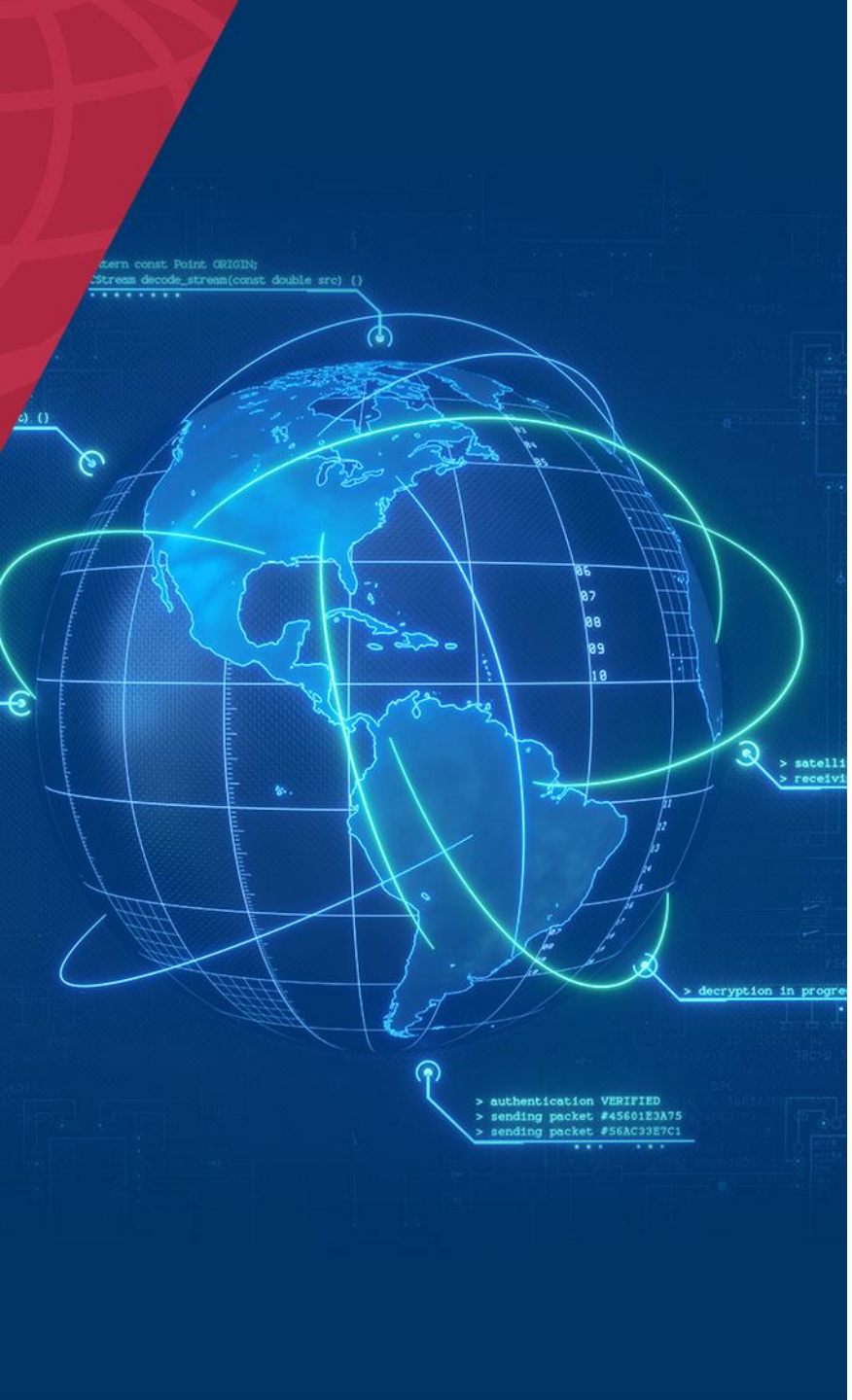
Categories – Additional Storage Data

- Name
- Storage Type
  - File
  - Object
  - Disk
  - Snapshot
- Disk Type
  - SSD
  - HDD
- Prices
  - Monthly
  - TB/Month
  - GB/Month
- Size
- Provider
- Continent
- Region
- Location (City)
- Date

Name	Storage Type	Disk Type	Storage Size (GB)	On Demand Monthly Price ($/Month)	On Demand Monthly Price ($/GB/Month)	Provider	Continent
P4	Disk	SSD	34.35968	5.28	0.153668485853186	Azure	North America
P6	Disk	SSD	68.71936	10.21	0.14857530687131	Azure	North America
P10	Disk	SSD	137.43872	19.71	0.143409368189692	Azure	North America
P15	Disk	SSD	274.87744	38.02	0.138316189207816	Azure	North America
P20	Disk	SSD	549.75488	73.22	0.133186630376069	Azure	North America
P30	Disk	SSD	1099.51	135.17	0.122936580840556	Azure	North America

Updated Analysis Process

- Analysis done with multiple tools (Python, Excel, TrueFindings®)
- Training and Test Sets
  - Training set: a subset to train a model (80% of data)
  - Test set: a subset to test the trained model (20% of data)
- Steps:
  1. Collected random sample of datapoints to split data into Training/Test Sets
  2. Develop models using Training Set
     a) *Piecewise Linear Regression (Virtual Machine Data)*
     b) *Benchmark Averages (Storage)*
  3. Validate the Test set against Training set model



Experimental Approach

Part 2: Develop Virtual Machine Pricing CERs

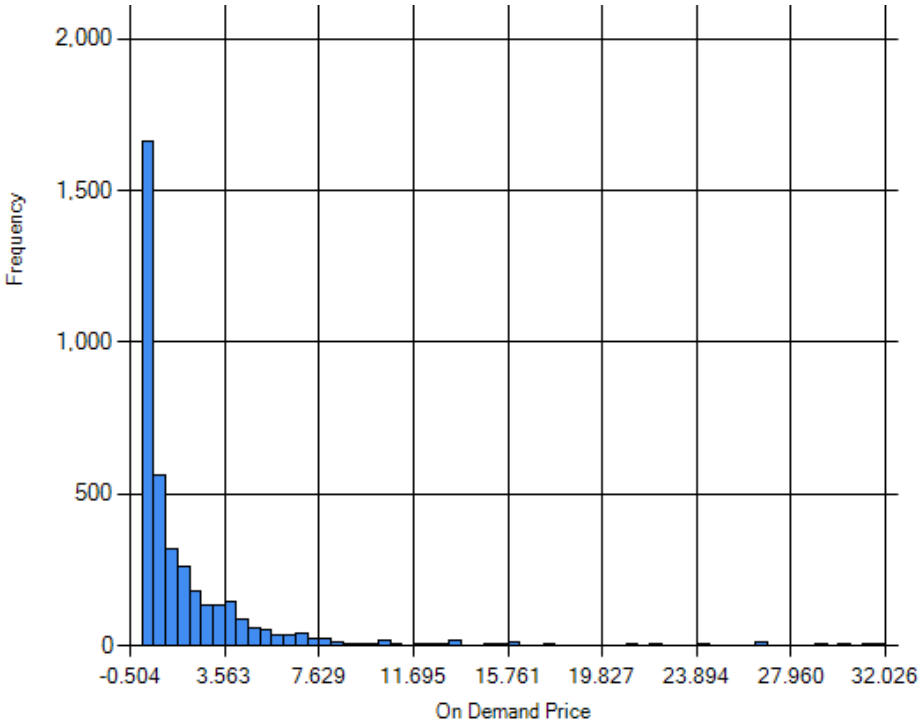
Data for Analysis

Assumptions:

- US locations only
- Type: Compute (ignored Kubernetes, which are price duplicates)
- Zero GPUs
- Non-zero values for CPUs, Memory, Price

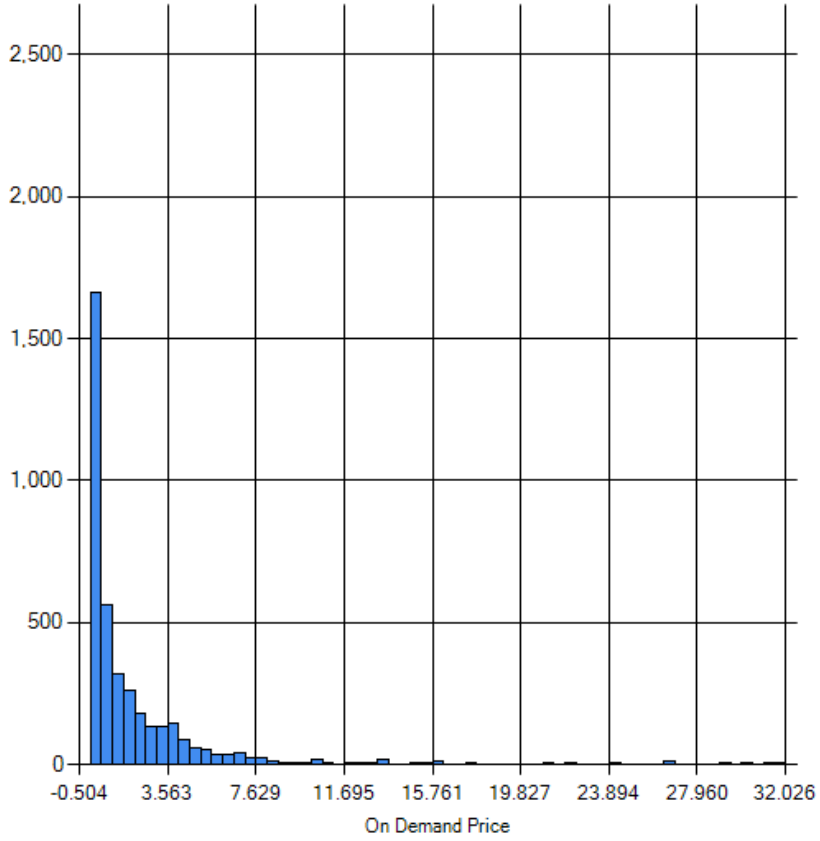
Value for On Demand Price ($/Hour)	
Count	3892
Min	0.005
Max	32.026
Mean	1.927
Median	0.745
Standard Deviation	3.364
Coefficient of Variation	1.745

Histogram – On Demand Price

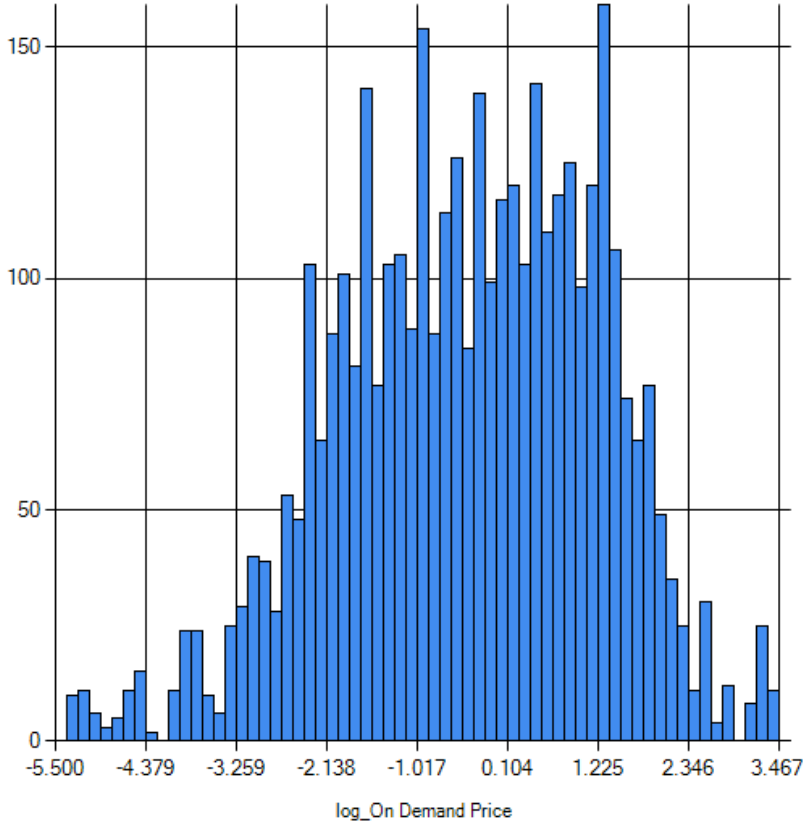


Data After Log Transform – ex. On Demand Pricing

Histogram – On Demand Price



Histogram – ln_On Demand Price



Training/Testing Split

- Randomly generated Training/Testing Data from datapoints
  - Only performed once - all Reserved Pricing data is a subset of On Demand Pricing

- Final counts for each pricing type (non-zero values):

Pricing Type	Training Set Size (80%)	Testing Set Size (20%)	Total Datapoints
On Demand	3113	779	**3892**
1 Year Reserved	2360	604	**2964**
3 Year Reserved	2237	560	**2797**

Model Building – Correlation

- Compared variables before and after Log Normal transform
  - Prices and log transformed prices had better correlation with respective original and log-transformed variables (e.g. ln_On Demand price correlated better with ln_Memory than Memory)

- Memory and CPUs may be colinear, but not an issue[1]

Variable 1	Variable 2	R Value
On Demand Price	Memory	0.943
On Demand Price	CPUs	0.813
On Demand Price	Local Disk Storage	0.275
CPUs	Memory	0.717
ln_On Demand Price	ln_Memory	0.967
ln_On Demand Price	ln_CPUs	0.939
ln_On Demand Price	ln_Local Disk Storage	0.410
ln_Memory	ln_CPUs	0.910

[1]Source: http://www.stat.tamu.edu/~hart/652/collinear.pdf

CER Development – First Steps

- Goal: produce Mean Absolute Percent Error (MAPE) to be within at least 30% for each category of pricing

- Initial analysis
  - Single and multivariate analysis with CPUs and Memory produced poor results with the test sets
  - VM instances with lower CPUs (less than 16) were *very* poorly predicted by these initial CERs

- Solution: split CERs by CPU amounts

CER Development – CPU Piecewise Regression

- Tested different boundary points using Python:
  - 4, 8, 16, 32, 48, 64 CPUs
  - Tested for both original and log normally transformed price
  - Optimized solution produces best MAPE and training/test sizes

- Every pricing type has 2 CERs
  a) <= 32 CPUs
  b) > 32 CPUs

CPU Range	Price Type	Training Set Size	Test Set Size
=< 32	On Demand	2441	611
> 32	On Demand	672	168
=< 32	1 Year Reserved	1791	459
> 32	1 Year Reserved	569	145
=< 32	3 Year Reserved	1676	422
> 32	3 Year Reserved	561	138

Final CERs

CPU Range	Formula
=< 32	*On Demand Price =* *Exp(0.444 * [ln_CPUs] + 0.019 * [ln_Local Disk] + 0.590 * [ln_Memory] - 3.787)*
> 32	*On Demand Price =* *0.037 * [CPUs] + 0.00007 * [Local Disk Storage] + 0.005 * [Memory] - 0.366*
=< 32	*1 Year Reserved Price =* *Exp(0.452 * [ln_CPUs] + 0.053 * [ln_Local Disk Storage] + 0.548 * [ln_Memory] - 4.331)*
> 32	*1 Year Reserved Price =* *0.022 * [CPUs] + 0.00005 * [Local Disk Storage] + 0.003 * [Memory] - 0.178*
=< 32	*3 Year Reserved Price =* *Exp(0.476 * [ln_CPUs] + 0.035 * [ln_Local Disk Storage] + 0.539 * [ln_Memory] - 4.723)*
> 32	*3 Year Reserved Price =* *0.015 * [CPUs] + 0.00003 * [Local Disk Storage] + 0.001 * [Memory] + 0.103*

Results: All Test Data

CPU Range	Formula	Mean % Error	MAPE
=< 32	On Demand Price*	-6.3%	30.0%
> 32	On Demand Price	-2.6%	15.0%
=< 32	1 Year Reserved Price*	-5.3%	25.9%
> 32	1 Year Reserved Price	-3.0%	15.1%
=< 32	3 Year Reserved Price*	-6.5%	24.5%
> 32	3 Year Reserved Price	-6.6%	15.2%

Regressions for CPUs > 32 performed significantly better than

<= 32 CPUs (*log transform equations), but met goal of <= 30% MAPE

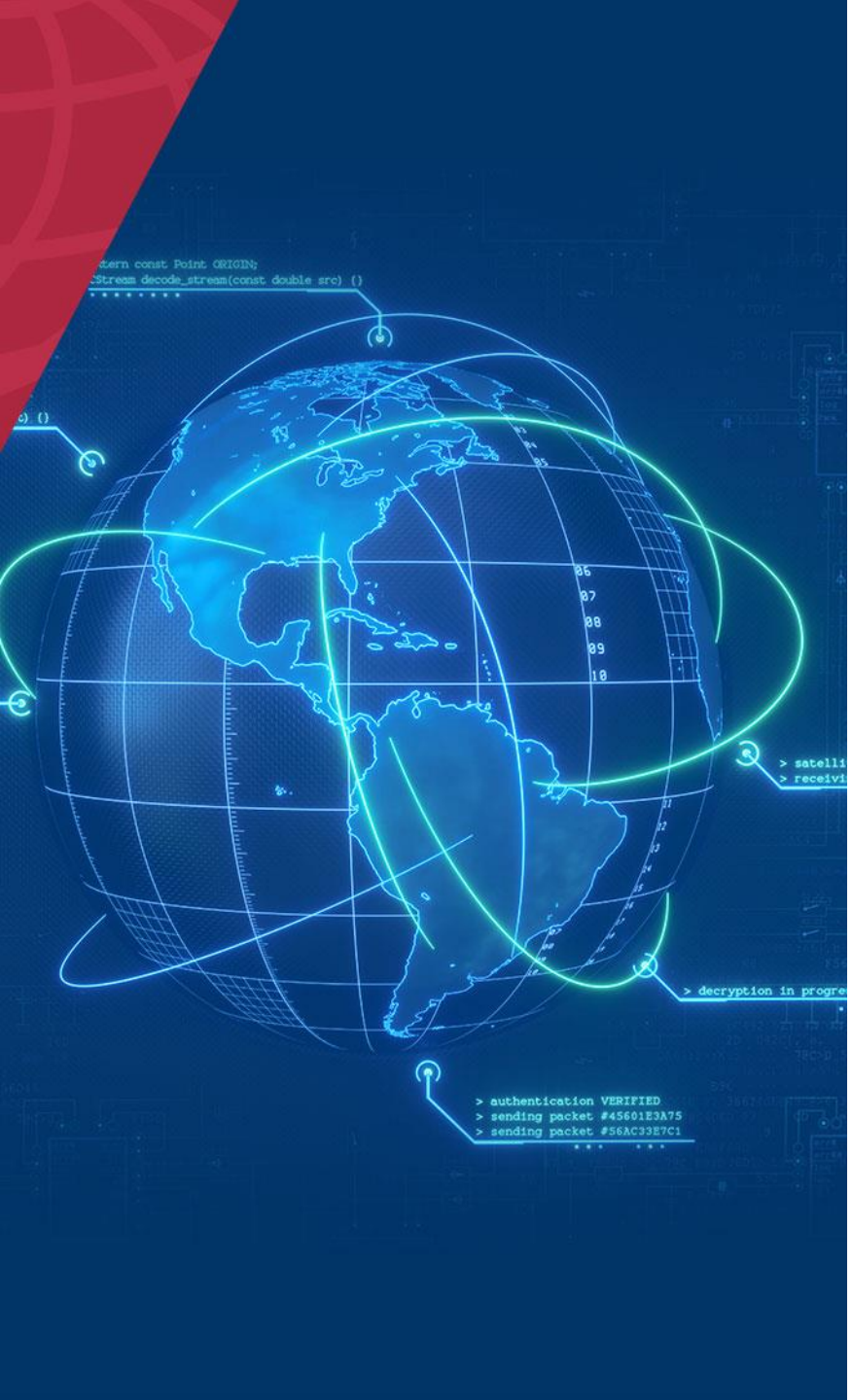
Results: Gov Test Data Comparison

Gov Test Data	On Demand	1 Yr Reserved	3 Yr Reserved
MAPE	**29.7%**	**27.2%**	**26.2%**
Number of Datapoints	108	97	87

All Test Data	On Demand	1 Yr Reserved	3 Yr Reserved
MAPE	**26.7%**	**23.2%**	**22.1%**
Number of Datapoints	780	604	560

Government VM Test data is a subset of initial data (Azure and Amazon only)

Gov results skewed by about 5 micro/nano VMs with poor predictions (100%+ error). Reserved CERs perform slightly better than On Demand.



Experimental Approach

Part 3: Develop Additional Storage Pricing Models*

*All storage prices are standard rates

Object Storage – Pricing Summary

Data Transfers, Operational Requests, Object Storage

Data Storage Prices (Per TB)	Avg Rate (Vendor Agnostic)
First 50 TB	22.45
Next 450	22.15
500+ TB	21.51
Data Transfer Out Prices (Per TB)	
0-1 TB	96.26
1-10 TB	93.70
10-50 TB	82.94
50-150 TB	71.68
150+ TB	57.09
Operational Request Prices (per 10k operations)	
WRITE Requests	0.0368
READ Requests	0.0034

Research Source: https://www.enterprisestorageforum.com/cloud-storage/cloud-storage-pricing.html

Block Storage – Pricing Summary

Snapshots, Disks (Block Storage) – Database averages

Disk Model

Disk Type	$/GB of Storage
HDD	$ 0.0513
SSD	$ 0.1019

Snapshot Model

$/GB of Snapshots
$ 0.0365

File Storage – Pricing Summary

Two types of pricing

- Performance: pay as you go storage
- Capacity: paying for a set amount of storage

Type of Pricing	Amazon	Azure	Google	IBM	Oracle	Alibaba
Performance ($/TB)	314.88	-	-	207.36	307.20	-
Capacity ($/TB)	-	284.64	222.35	-	-	271.93

Average Monthly Rate = $264.87/TB

Object Storage Results

	# of Training datapoints	# of Test datapoints	Abs Avg % Error
Data Storage	60	15	**10.30%**
Data Transfer	65	15	**13.18%**
Operational Requests	42	8	**16.36%**
MAPE			**12.71%**

File Storage & Disk Storage Results

- File Storage

File Storage	# of Training datapoints	# of Test datapoints	Abs Avg % Error
File	34	8	**19.29%**

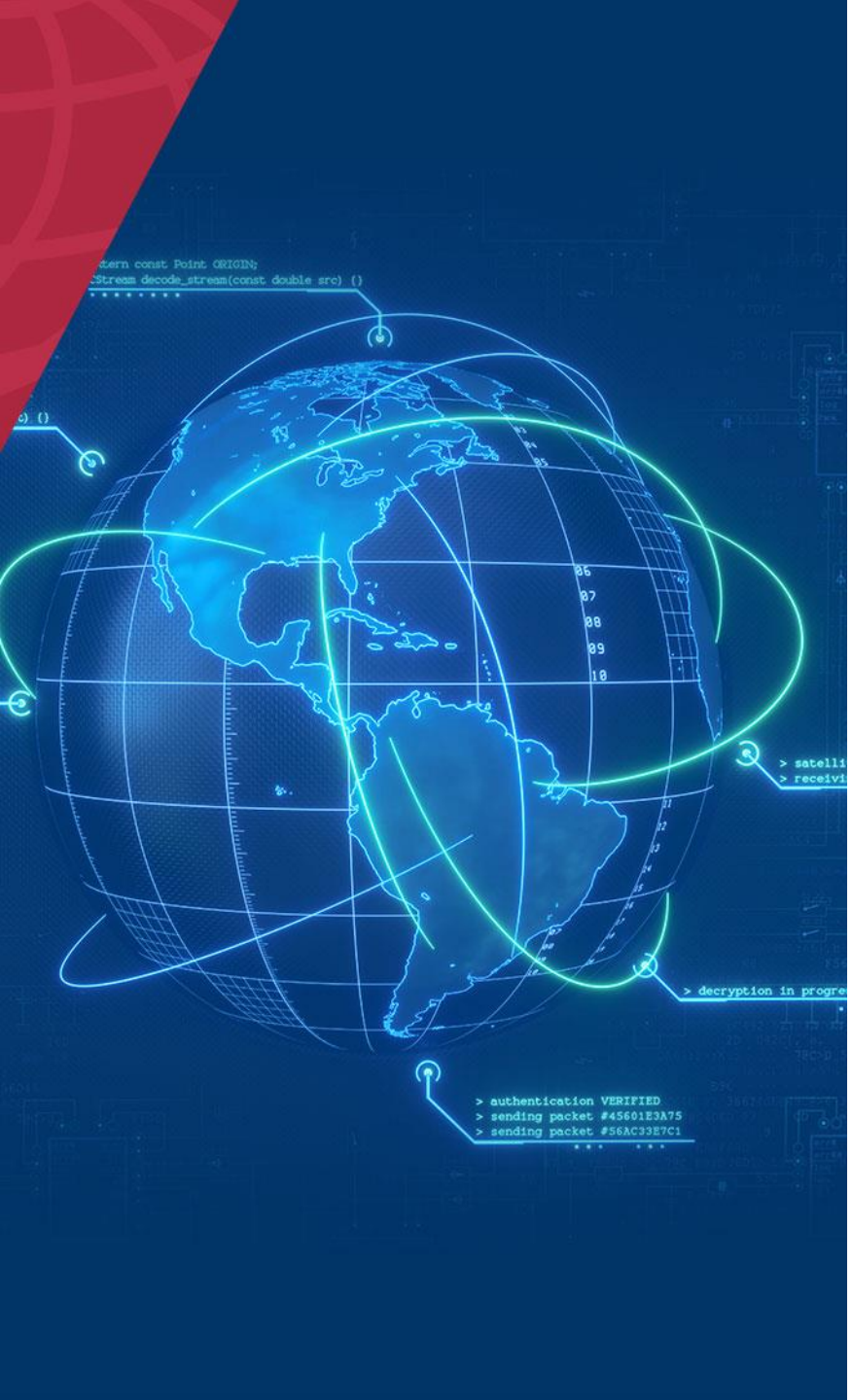
- Disk Storage
  - Government price datapoints were included in the dataset before the initial test/train split

Disk Storage	# of Training datapoints	# of Test datapoints	Abs Avg % Error
SSD	263	66	**32.51%**
HDD	144	36	**27.27%**
MAPE			**30.66%**

Government Storage Price Validation

- Government price datapoints were collected for US Locations where available
- Validation of government prices against model:

Storage Type	Abs Avg % Error
Object Storage	32.30%
File Storage	29.37%
Disk Storage - SSD	30.90%
Disk Storage - HDD	34.91%



Summary

Overall Results

- Models produce marginally acceptable results for Government pricing (around 30% MAPE)
- VMs
  - Prices can be modeled as a function of Memory, CPUs in piecewise regression with some Log Normal transforms
  - CER for instances with more than 32 CPUs have more predictive accuracy than <= 32 CPUs
- Storage
  - Benchmark models are adequate
  - Object and File perform better than Disk storage
  - Snapshot dataset is too small to perform validation
  - Training/Test datasets may be too small in some cases

Conclusions

- CERs/models developed from multiple commercial provider rates able to predict costs for:
  - Compute Environment (servers, RAM)
  - Storage Environment (capacity)

- Open source database can help guide user to pick prices and form CERs/models based on specific needs

Next Steps – Current Models

- Storage
  - Backup costs
- Virtual Machines
  - Default technical specifications for calculator (RAM, # of CPUs, etc.)
- Expand research to predict other cloud-related costs
  - Other cloud service costs (if data is easily obtainable)
  - Transition/migration costs

Points of Contact

Cara Cuiule

Cost Research Analyst

PRICE Systems

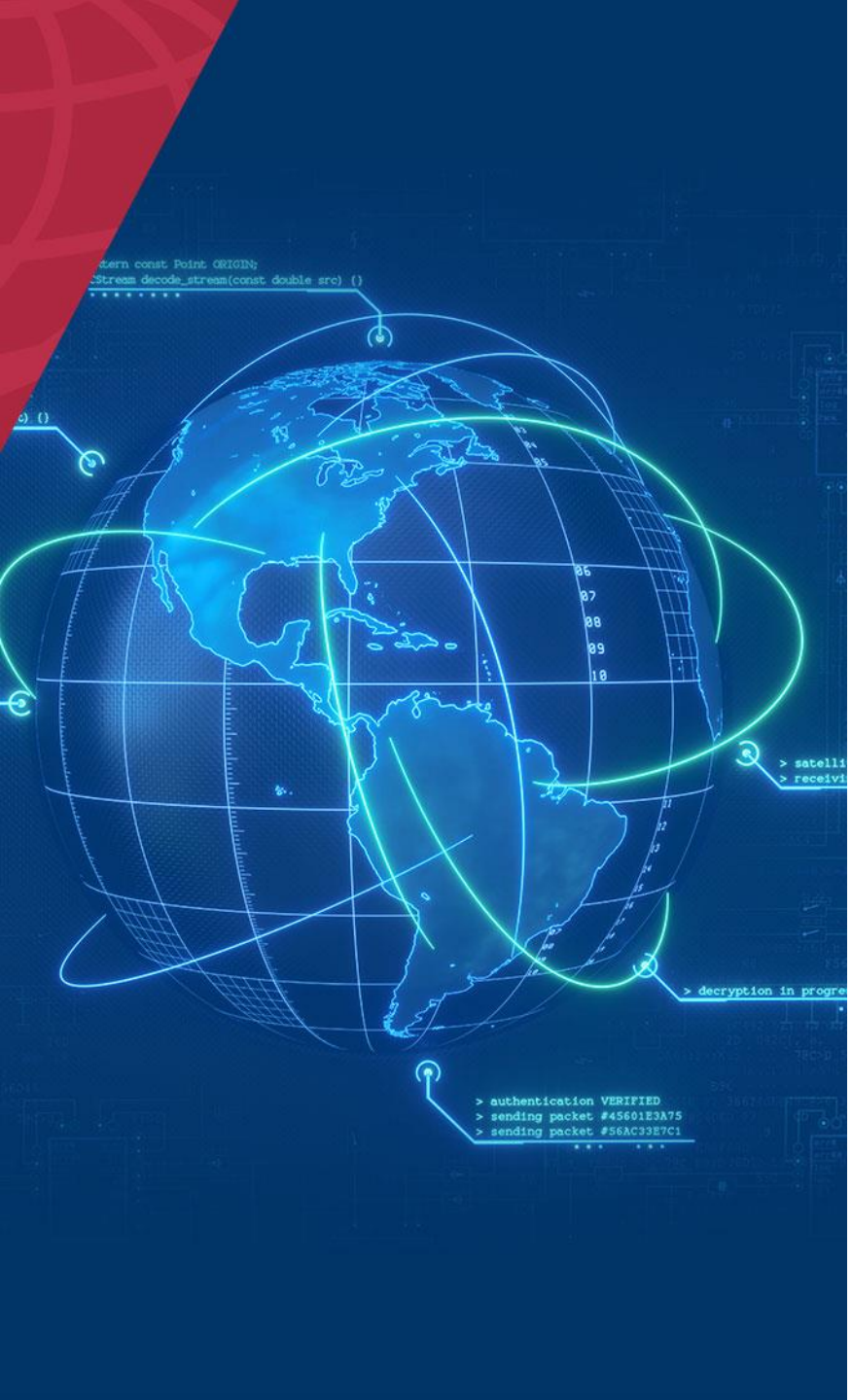
Cara.Cuiule@pricesystems.com

Amanda Ferraro

Cost Research Analyst

PRICE Systems

Amanda.Ferraro@pricesystems.com



Backup

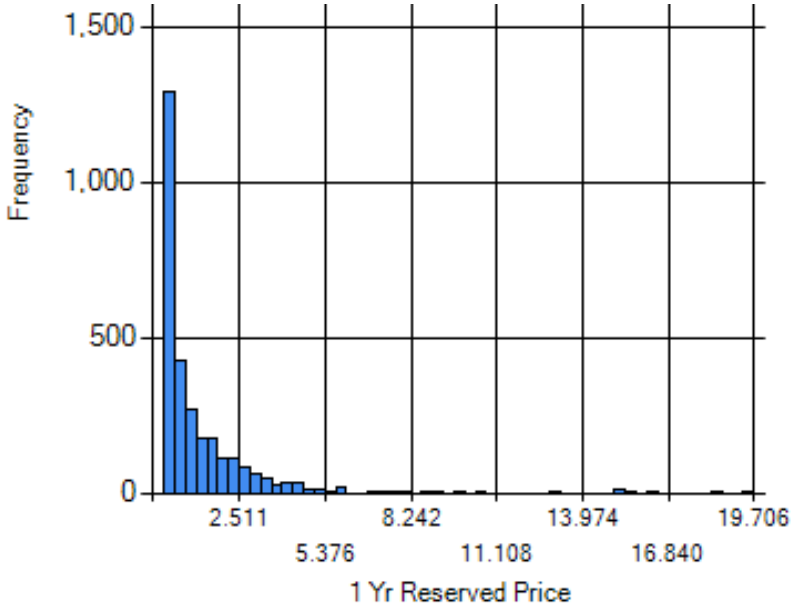
Data Collection Challenges

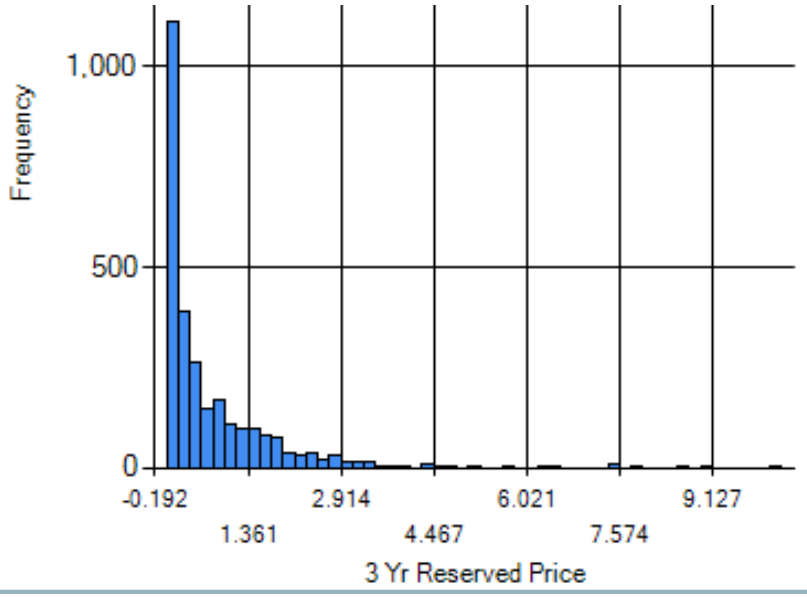
- Normalizing data
- Partially automated process
- Data not included in the current research (due to time constraints):
  - Google reserved prices (only provider missing from reserved)
  - Some Govt Pricing (not publicly available)
  - Graphic Processing Unit (GPU) properties

Additional Data Profiling- USA VM Prices

Data Element	Min Value	Max Value
Number of CPUs	1	208
Memory (RAM) (GB)	0.5	4,191
Local Disk (GB)	0	64,424
1 Year Reserved Price* ($/Hr)	0.003	19.706
3 Year Reserved Price* ($/Hr)	0.002	10.292

*Non-zero values only, see slide 22 for datapoint counts





VMs - Reserved Prices Correlation

Variable 1	Variable 2	R Value
1 Yr Reserved Price	Memory	0.945
1 Yr Reserved Price	CPUs	0.817
1 Yr Reserved Price	Local Disk Storage	0.313
1 Yr Reserved Price	Network Performance	0.255
ln_1 Yr Reserved Price	ln_Memory	0.967
ln_1 Yr Reserved Price	ln_CPUs	0.944
ln_1 Yr Reserved Price	ln_Local Disk Storage	0.452
ln_Memory	ln_CPUs	0.912
3 Yr Reserved Price	Memory	0.893
3 Yr Reserved Price	CPUs	0.857
3 Yr Reserved Price	Local Disk Storage	0.379
3 Yr Reserved Price	Network Performance	0.345
ln_3 Yr Reserved Price	ln_Memory	0.968
ln_3 Yr Reserved Price	ln_CPUs	0.952
ln_3 Yr Reserved Price	ln_Local Disk Storage	0.443
ln_Memory	ln_CPUs	0.914

VMs - About the Local Disk Variable

- CERs without Local Disk Storage overall had slightly worse results

CPU Range	Formula	No Local Disk CER MAPE	Local Disk CER MAPE
=< 32	On Demand Price*	29.9%	30.0%
> 32	On Demand Price	16.6%	15.0%
=< 32	1 Year Reserved Price*	26.9%	25.9%
> 32	1 Year Reserved Price	16.8%	15.1%
=< 32	3 Year Reserved Price*	25.2%	24.5%
> 32	3 Year Reserved Price	17.9%	15.2%

Object Storage – Pricing Summary

Data Transfers, Operational Requests, Object Storage

	Amazon	Microsoft	Google	IBM	Oracle	Alibaba	
Data Storage Prices (Per TB)							**Avg Rate (Vendor Agnostic)**
First 50 TB	24.32	19.7632	22.23543	22.528	26.112	19.712	**22.45**
Next 450	23.296	19.008	22.23543	22.528	26.112	19.712	**22.15**
500+ TB	22.272	18.2528	22.23543	20.48	26.112	19.712	**21.51**
Data Transfer Out Prices (Per TB)							
0-1 TB	92.160		122.88	92.16		77.824	**96.26**
1-10 TB	92.160		112.64	92.16		77.824	**93.70**
10-50 TB	87.040		81.92	92.16		70.656	**82.94**
50-150 TB	71.680		81.92	71.68		61.44	**71.68**
150+ TB	51.200		81.92	51.2		44.032	**57.09**
Operational Request Prices (per 10k operations)							
WRITE Requests	0.051	0.051	0.05	0.05	0.003	0.016	**0.0368**
READ Requests	0.004	0.004	0.004	0.004	0.003	0.001	**0.0034**

Research Source: https://www.enterprisestorageforum.com/cloud-storage/cloud-storage-pricing.html