Is capacity management needed in the cloud?

Kevin McLaughlin

Capital One

- **Millions of customer accounts**
- **One of the largest digital banks**
- **~20 years old**

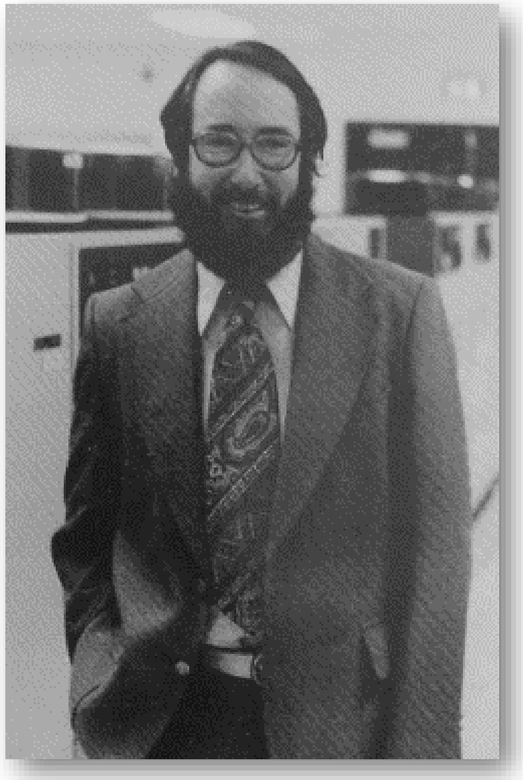
tl;dr

Yes, but it looks vastly different from traditional capacity management

The mission of capacity management is to ensure sufficient processing power exists to support current and planned business needs

"…manufacturers market combinations of hardware and software that provide general-purpose data processing. [Their] goal is to maximize [their] profit while providing sufficient hardware and software to meet the customer's needs. The customer, however, wants to minimize the cost of the hardware and software while meeting users' requirements."

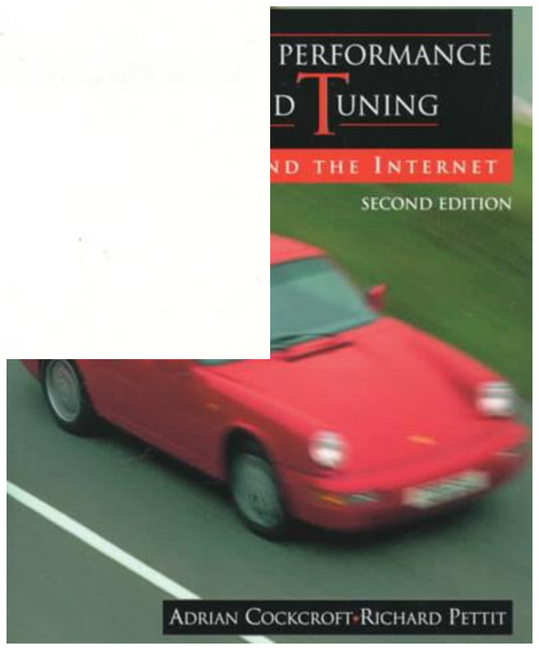
-Barry Merrill, "Merrill's Guide to Computer Performance Evaluation" 1980



Sun Performance and Tuning

May your Sun's perform in time!
this book is obsolete!
Adrian Cockcroft

Adrian Cockcroft



FFIEC
IT Examination HandBook InfoBase

IT Booklets | Resources | Reference Materials | Presentations | Glossary | Help | Search | What's New

- Audit
- Business Continuity Planning
- Development and Acquisition
- E-Banking
- Information Security
- Management
  - Introduction
  - Risk Overview
  - Roles and Responsibilities
  - IT Risk Management Process
  - Management Considerations for Technology
  - Appendix A: Examination Procedures
  - Appendix B: Laws, Regulations, and Guidance

Welcome » IT Booklets » Operations » Risk Monitoring and Reporting » Capacity Planning

Capacity Planning

Capacity planning involves the use of baseline performance data to model and project future needs. Capacity planning should address internal factors (growth, mergers, acquisitions, new product lines, and the implementation of new technologies) and external factors (shift in customer preferences, competitor capability, or regulatory or market requirements). Management should monitor technology resources for capacity planning including platform processing speed, core storage for each platform's central processing unit, data storage, and voice and data communication bandwidth.

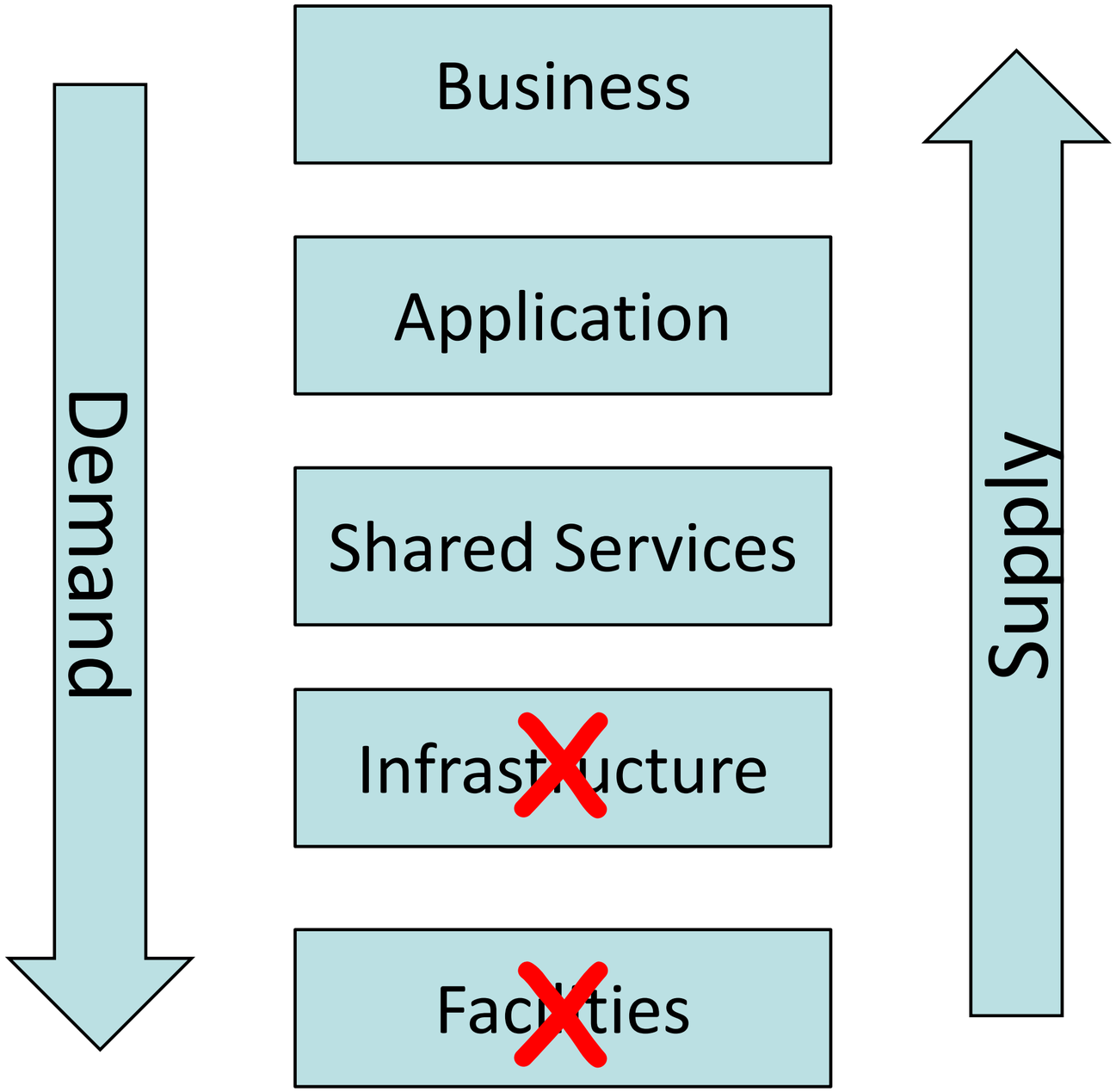
Capac
proces
staff s

Previo
Perfor

Objective 5: Determine whether there are adequate controls to manage the operations-related risks.

1. Determine whether management has implemented and effectively utilizes operational control programs, processes, and tools such as:

- Performance management and capacity planning;
- User support processes;
- Project, change, and patch management;
- Conversion management;
- Standardization of hardware, software, and their configuration;
- Logical and physical security;
- Imaging system controls;
- Environmental monitoring and controls; and
- Event/problem management.

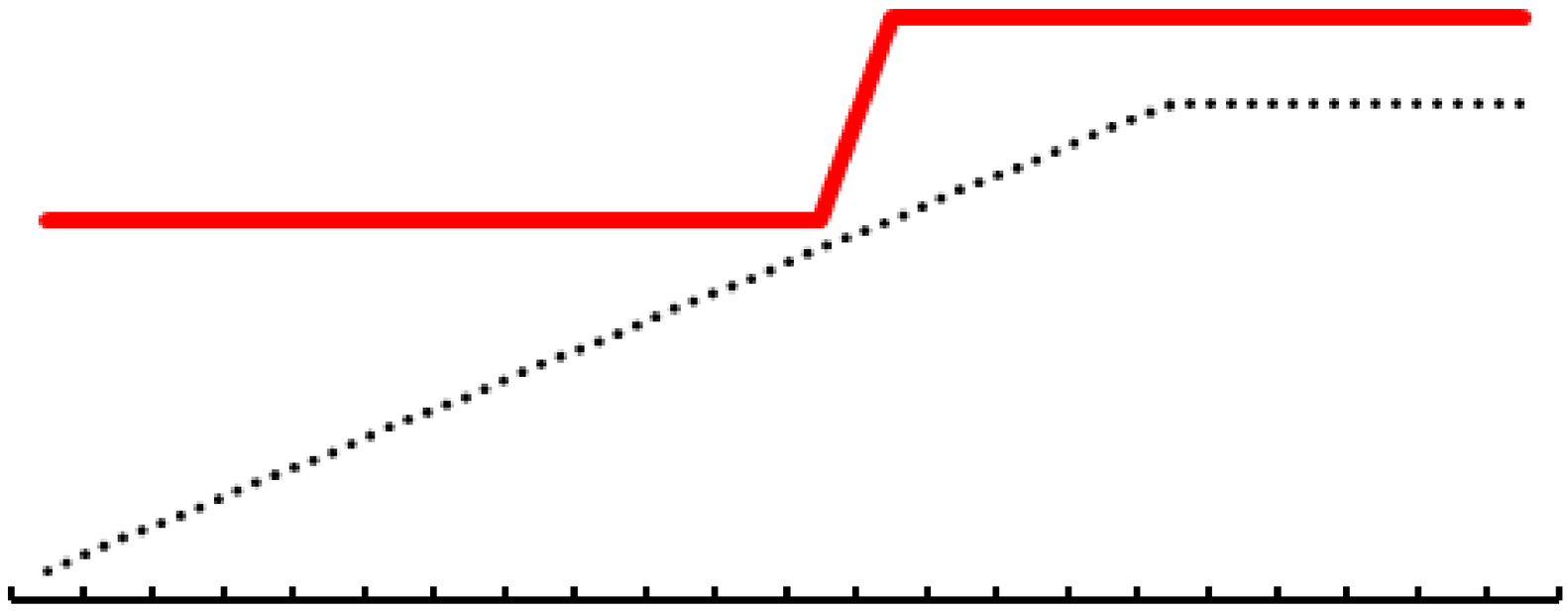


Source: https://www.cmg.org/wp-content/uploads/2013/10/CMG-Revolutionary_Capacity_Planning-PAPER.pdf

Layer	Description	Traditional Capacity Management	Public Cloud Capacity Management	
			IaaS (eg ec2)	PaaS (eg RDS)
Business Demand	Track and forecast business metrics and project related demand	✓	✓	✓
Application	Track transaction volumes and performance metrics for workload characterization	✓	✓	✓
Shared Services	Ensure sufficient capacity exists for shared services	✓	✓	✗
Infrastructure - Instances	Track and forecast for cost and efficiency recommendations	✓	✓	✗
Infrastructure - hosts	Track and forecast for cost, hardware acquisition recommendations	✓	✗	✗
Data Center	Track and forecast power, floorspace, cooling for expansion recommendations	✓	✗	✗

Traditional	Public Cloud
Reclamation is hard	Reclamation is possible, and can be semi- or fully automated
Hardware matters	Hardware can be ignored (mostly)
Horizontal scaling is hard	Horizontal scaling can be easily automated (Auto Scaling)
Vertical scaling is hard, but capped only by $	Vertical scaling is manual (sort of), and capped by cloud provider offerings
Data for analysis is hard to get	Data for analysis is easy to get (Cloudwatch)
Poor performance and/or Insufficient hardware impacts business	Poor performance impacts business
Need to predict consumption for budgeting	Need to predict consumption for budgeting
Focus on application level limits	Focus on VPC level limits

Even cloud providers have limits



Amazon Elastic Compute Cloud (Amazon EC2) Limits

Resource	Default Limit
Elastic IP addresses for EC2-Classic	5
Security groups for EC2-Classic per instance	500
Rules per security group for EC2-Classic	100
Key pairs	5,000
Throttle on the emails that can be sent from your Amazon EC2 account	Throttle applied
On-Demand instances	Limits vary depending on instance type. For more information, see How many instances can I run in Amazon EC2.
Spot Instances	Limits vary depending on instance type, region, and account. For more information, see Spot Instance Limits.
Reserved Instances	20 instance reservations per Availability Zone, per month.
Dedicated Hosts	Up to 2 Dedicated Hosts per instance family, per region can be allocated.
AMI Copies	Destination regions are limited to 50 concurrent AMI copies at a time, with no more than 25 of those coming from a single source region.

Source: http://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html

On this page:

Auto Scaling

- # Groups
- Launch Configs

EBS

- Active Volumes
- Provisioned IOPS
- Magnetic Volume Storage (GB)
- Active Snapshots
- Gen’l Purpose Volume Storage

EC2

- Reserved Instances
- On-Demand Instances

VPC

- # VPCs
- Internet Gateways

ELB

- Active Load Balancers

RDS

- # DB Instances
- DB Parameter Groups
- DB Security Groups
- Storage Quota
- DB Snapshots per user

Troubleshooting Instance Capacity

The following errors are related to instance capacity.

Error: InsufficientInstanceCapacity

If you get an `InsufficientInstanceCapacity` error when you try to launch an instance, AWS does not currently have enough available capacity to service your request. Try the following:

- Wait a few minutes and then submit your request again; capacity can shift frequently.
- Submit a new request with a reduced number of instances. For example, if you're making a single request to launch 15 instances, try making 3 requests for 5 instances, or 15 requests for 1 instance instead.
- Submit a new request without specifying an Availability Zone.
- Submit a new request using a different instance type (which you can resize at a later stage). For more information, see Resizing Your Instance.
- Try purchasing Reserved Instances. Reserved Instances are a long-term capacity reservation. For more information, see: Amazon EC2 Reserved Instances.

Error: InstanceLimitExceeded

If you get an `InstanceLimitExceeded` error when you try to launch an instance, you have reached your concurrent running instance limit. For new AWS accounts, the default limit is 20. If you need additional running instances, complete the form at Request to Increase Amazon EC2 Instance Limit.

Source: http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-capacity.html

awslabs / **aws-limit-monitor**

Watch 7 | Star 14 | Fork 3

Code | Issues 0 | Pull requests 0 | Pulse | Graphs

Customizable Lambda functions to proactively notify you when you are about to hit an AWS service limit. Requires Enterprise or Business level support to access Support API.

8 commits | **1** branch | **0** releases | **3** contributors

Branch: **master** | **New pull request** | **Create new file** | **Upload files** | **Find file** | **Clone or download**

hyandell committed on GitHub Update NOTICE.txt — Latest commit 6260812 17 days ago

AWSLimits.cfn	Initial Push	2 months ago
LICENSE.txt	Create LICENSE.txt	18 days ago
NOTICE.txt	Update NOTICE.txt	17 days ago
README.md	Initial Push	2 months ago
configuration.py	added source headers	18 days ago
limitCheck.py	added source headers	18 days ago
limitMaster.py	added source headers	18 days ago
limits.zip	Initial Push	2 months ago

Source: https://github.com/awslabs/aws-limit-monitor

See also https://aws.amazon.com/about-aws/whats-new/2014/05/21/aws-trusted-advisor-now-monitors-service-limits-for-amazon-ses-amazon-vpc-and-auto-scaling/

Cloud Custodian

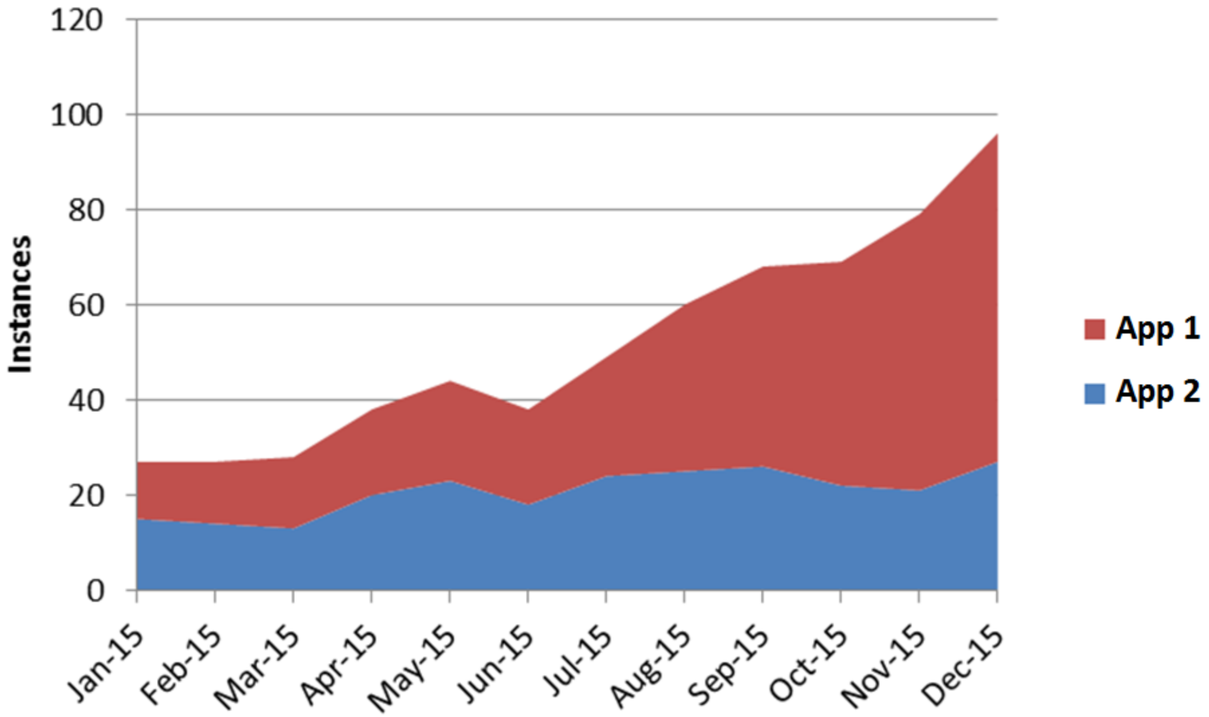
Cloud Custodian is a rules engine for AWS fleet management. It allows users to define policies to enable a well managed cloud infrastructure, that's both secure, and cost optimized. It consolidates many of the adhoc scripts organizations have into a lightweight and flexible tool, with unified metrics and reporting.

```
policies:
- name: remediate-extant-keys
  description: |
    Scan through all s3 buckets in an account and ensure all objects
    are encrypted (default to AES256).
  resource: s3
  actions:
    - encrypt-keys
```

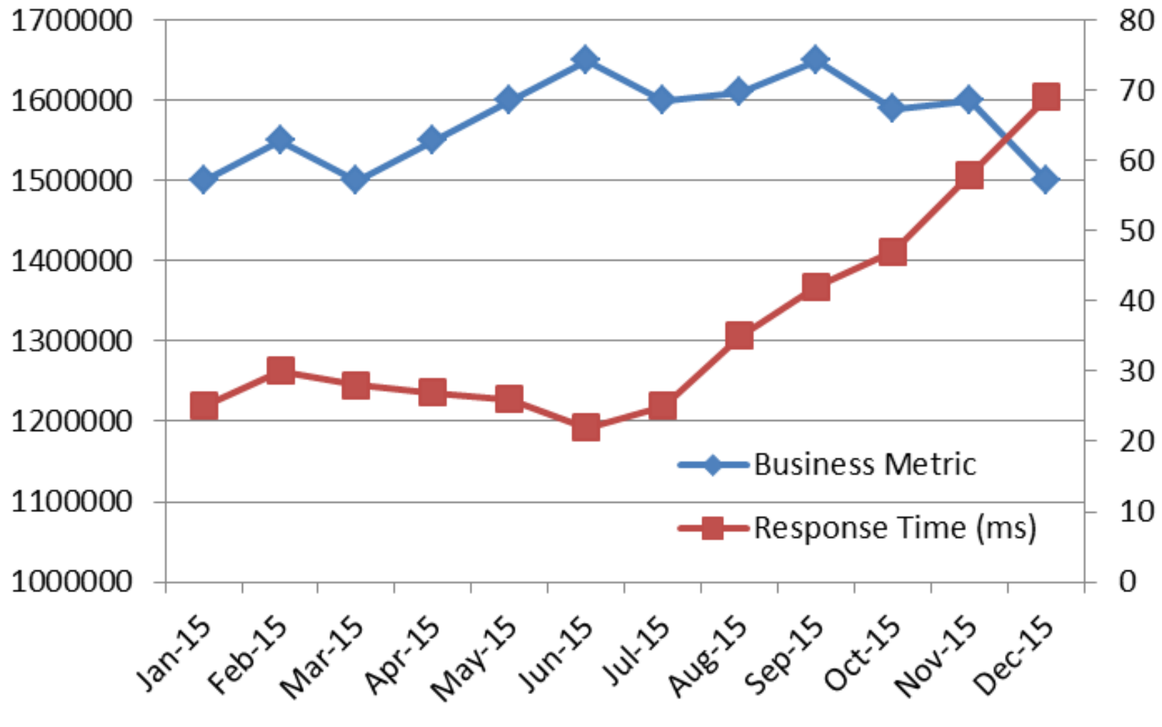
```
- name: ec2-require-non-public-and-encrypted-volumes
  resource: ec2
  description: |
    Provision a lambda and cloud watch event target
    that looks at all new instances and terminates those with
    unencrypted volumes.
  mode:
    type: cloudtrail
    events:
        - RunInstances
  filters:
    - type: ebs
      key: Encrypted
      value: false
  actions:
    - terminate
```

- Stolen CPU
- Start up times
- Predictive Auto Scaling (up and down)
- Instance counts
- Waste

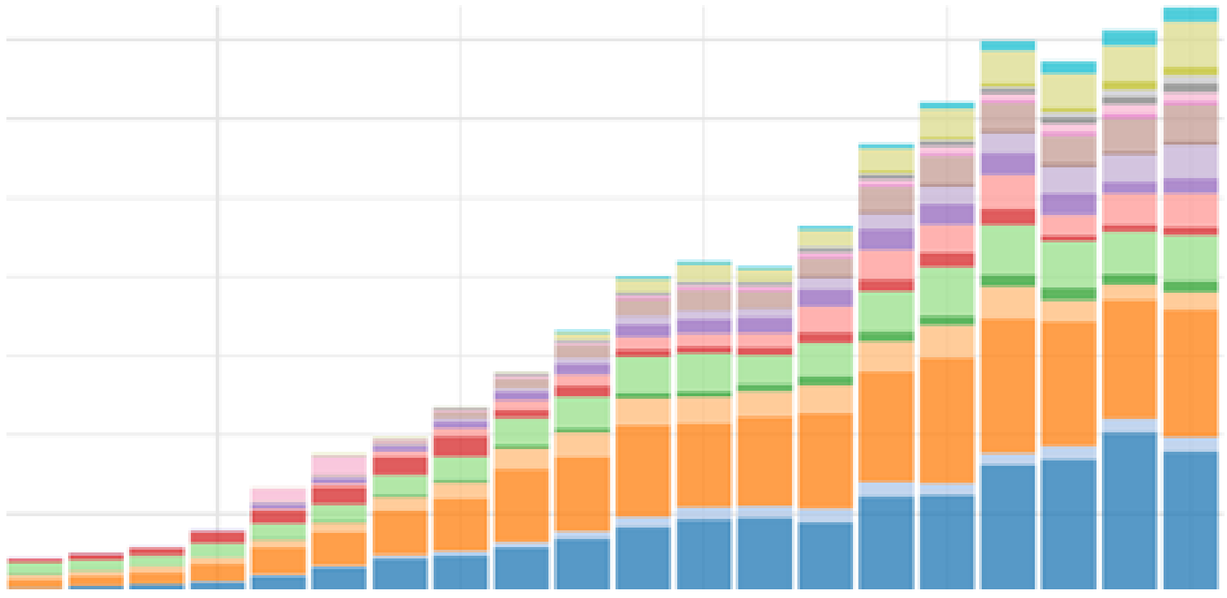
Change in ASG behavior

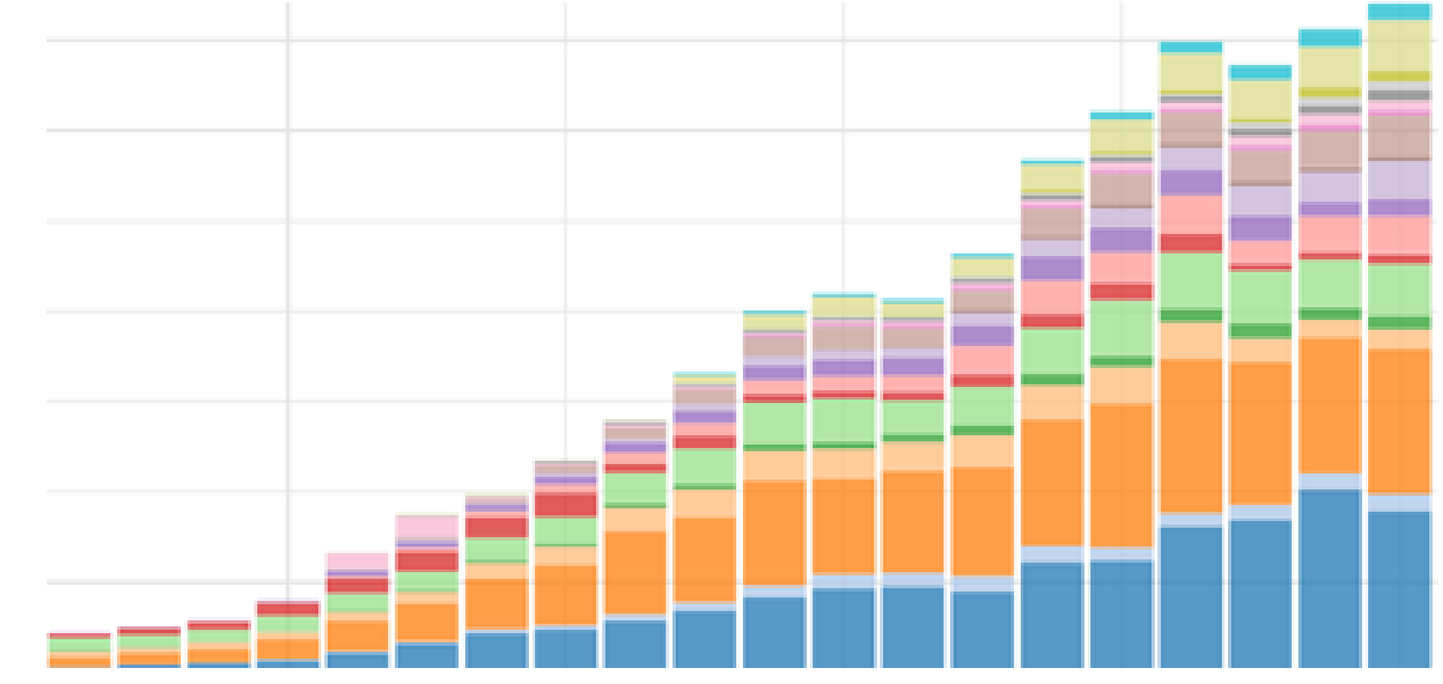


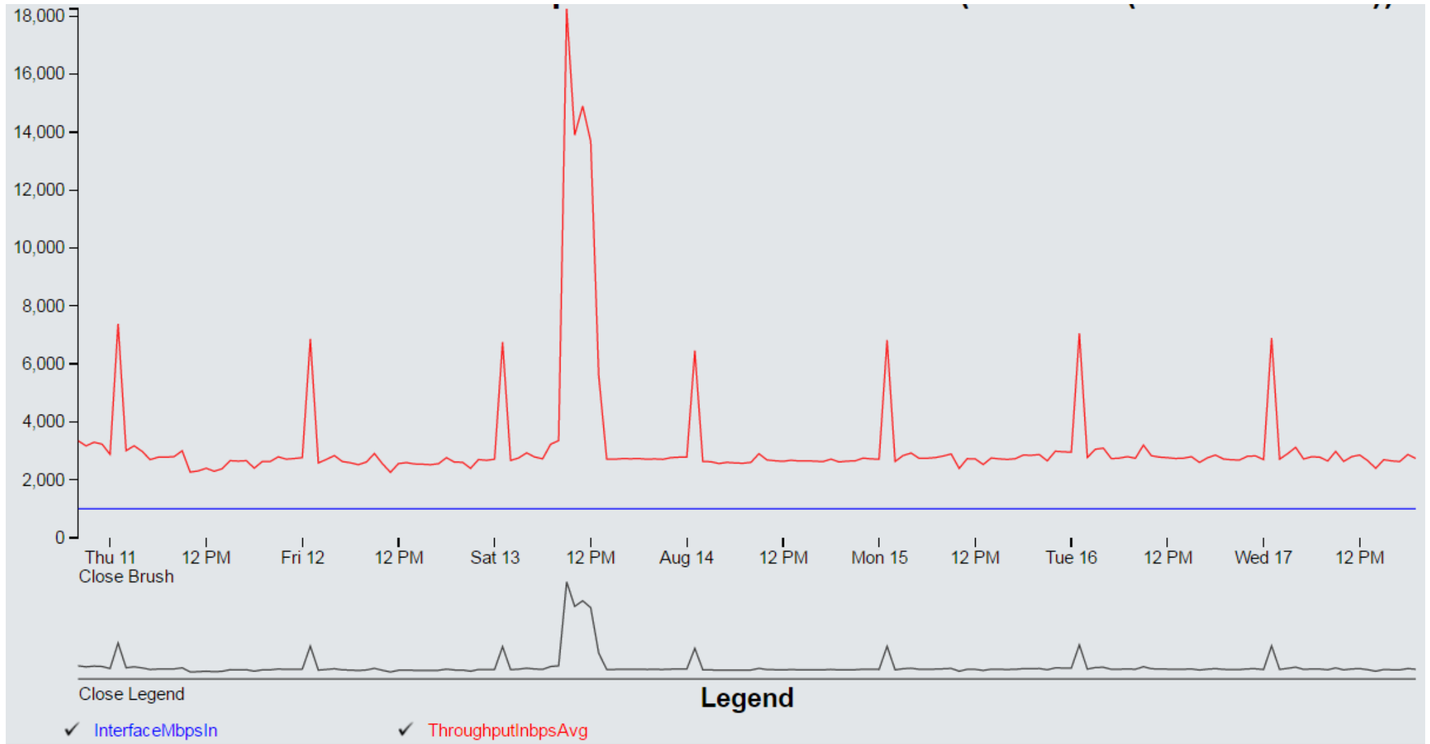
Unexplained performance degradation



Budgeting

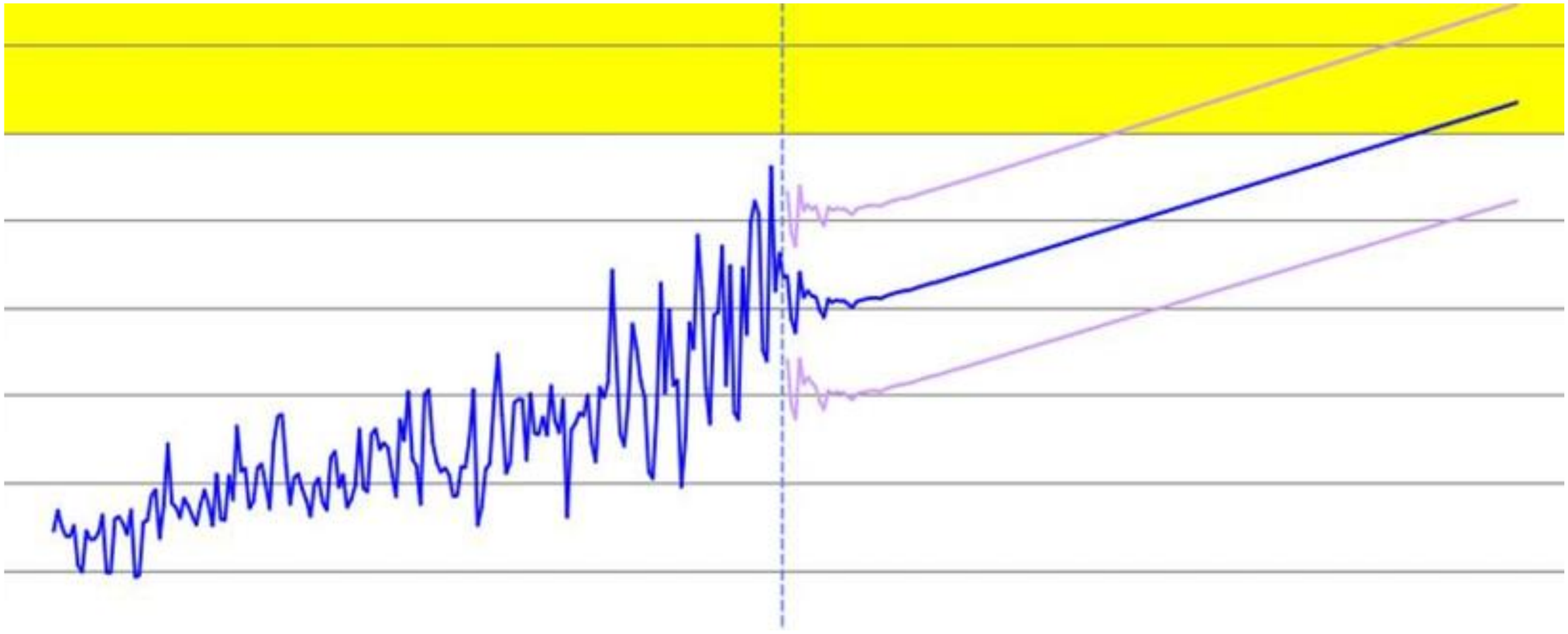





18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
Thu 11
12 PM
Fri 12
12 PM
Sat 13
12 PM
Aug 14
12 PM
Mon 15
12 PM
Tue 16
12 PM
Wed 17
12 PM
Close Brush
Close Legend
Legend
InterfaceMbpsIn
ThroughputInbpsAvg

	Capacity Management	Performance Management
Data Granularity	Medium	High
Time Window	Weeks to years (past, present, and future)	Milliseconds to hours (past and present)
Reporting Trigger	Calendar (mostly)	Event
Optimization Focus	Utilization (over and under)	Response Time/Throughput
Resource Focus	Prediction of needed	Allocation of existing
Skillset	Analyst + Tech	Engineer + Analyst

Getting to the cloud – watch your pipes



Skill	Traditional	Public Cloud
Analytics	Need to have	Need to have
Deep OS/Vendor expertise	Need to have	Nice to have
Knowledge of provisioning process	Need to have	Don’t need
Perceive infrastructure as ephemeral	Don’t need	Need to have
Coding	Specialized	General
Comfort with concept of something logical running on something physical	Need to have	Need to have

Is capacity management needed in the cloud?

- **Limits still exist**
- **Budgets still exist**
- **An outage due to a logical limit feels the same to the customer as an outage due to a physical limit**

ALL JOB OPENINGS

Manager - Cloud Capacity Engineering

Los Gatos, California

Netflix arguably runs one of the largest, most dynamic, and architecturally advanced infrastructure on the cloud today. Netflix engineers operate at maximum velocity with the ability to provision on-demand capacity for thousands of microservices at the push of a button, whether it be for five instances or 5,000. Your leadership in the capacity management space will allow us to efficiently scale our cloud footprint while maintaining our unparalleled innovation velocity.