

Behind the Curtain:

How we run Exchange Online

Jeff Mealiffe
Principal Program Manager
Microsoft

Setting some context...

all the things we can (or not) share

This is not a marketing or a sales session!

MY REPORT COMES TO THE CONCLUSION THAT CLOUD TECHNOLOGY IS OF NO USE TO THIS COMPANY. I'LL UPLOAD IT TO DROP BOX SO YOU CAN TAKE A LOOK AT IT.



© D. Fletcher for CloudTweaks.com

Not here to push anyone to the cloud—simply sharing the journey as we know you all love Exchange, technology and are keeping a watchful eye on where the cloud is heading

Good news is: we're going to show you even more about how Exchange Online runs at massive scale

Bad news is: yep, a lot of what we'll show you is engineering-cloud-stuff—not directly applicable to on-premises

Important caveats

We are going to try and show you details

Including technology which only lives @ Microsoft

Process and culture

Technology backing the service

BUT...

Some of these are outside the Exchange product, are only used within Microsoft

Some points will be necessarily vague or omitted altogether, e.g.

HW/Vendor: we cycle through a LOT of gear from a variety of vendors

Interdependent systems that do not exist in the enterprise product

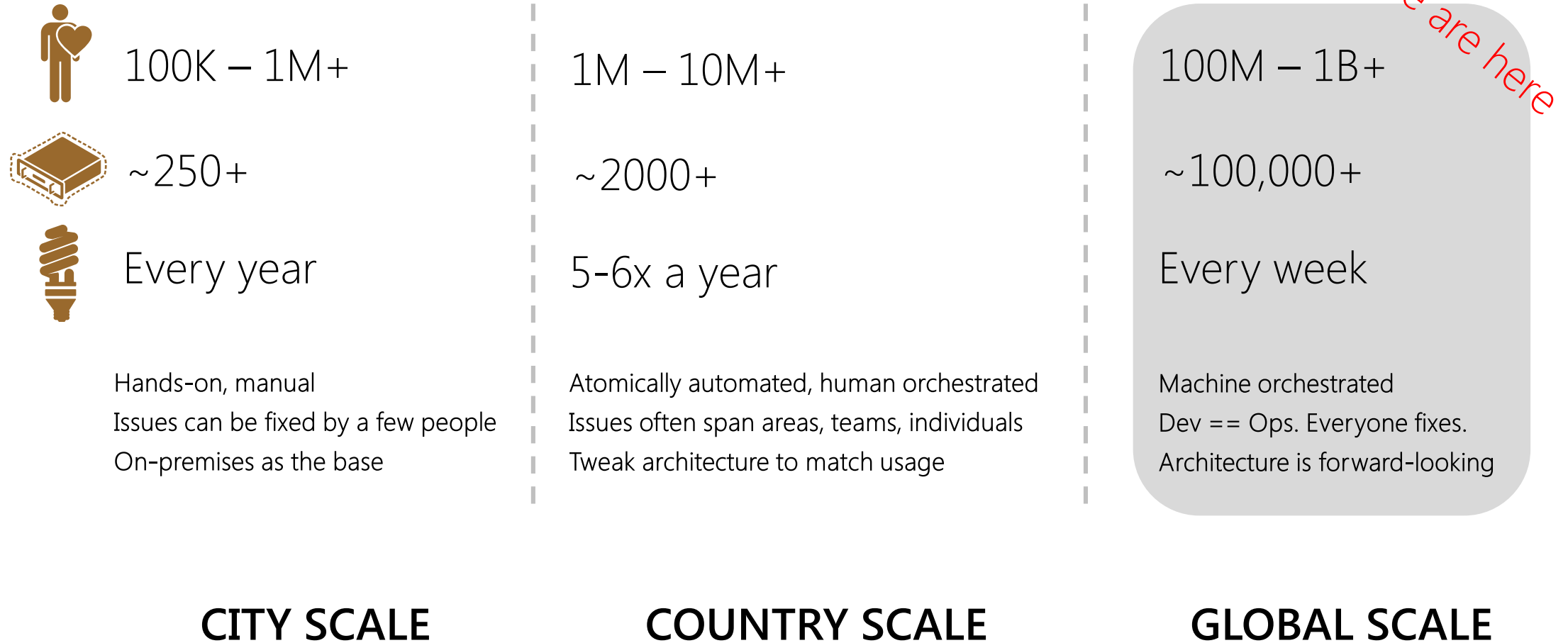
Exact numbers of things—always changing, easily misinterpreted

EXACT DETAILS NOT AVAILABLE

Exchange Online Scale

#'s, figures and all such things

The Evolution of Online



CITY SCALE

COUNTRY SCALE

GLOBAL SCALE

Everything changes at global scale

From

Turnkey software (SCOM, Windows, SQL)

Generalist workforce

Small failure domains

Errors that can be ignored or dealt with later

Easier to scale to customer issues

To

Highly customized, purpose driven automation

Specialists owning their piece of the puzzle e2e

Potentially HUGE failure domains

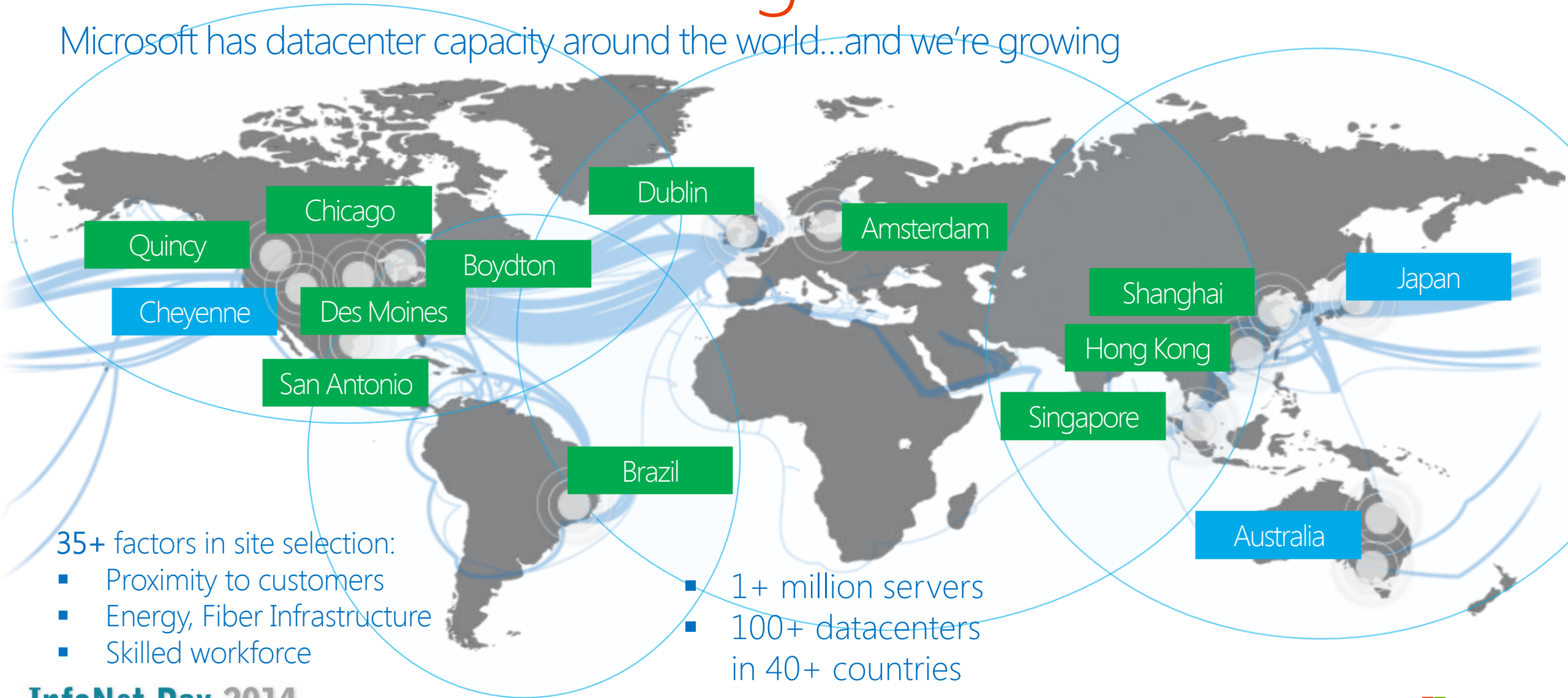
.01% error rate at 1B daily transactions is HUGE

Every action has consequences to vast number of users

Everything eventually breaks, we have to continuously (re)invent

Global Scale: DC growth

Microsoft has datacenter capacity around the world...and we're growing



35+ factors in site selection:

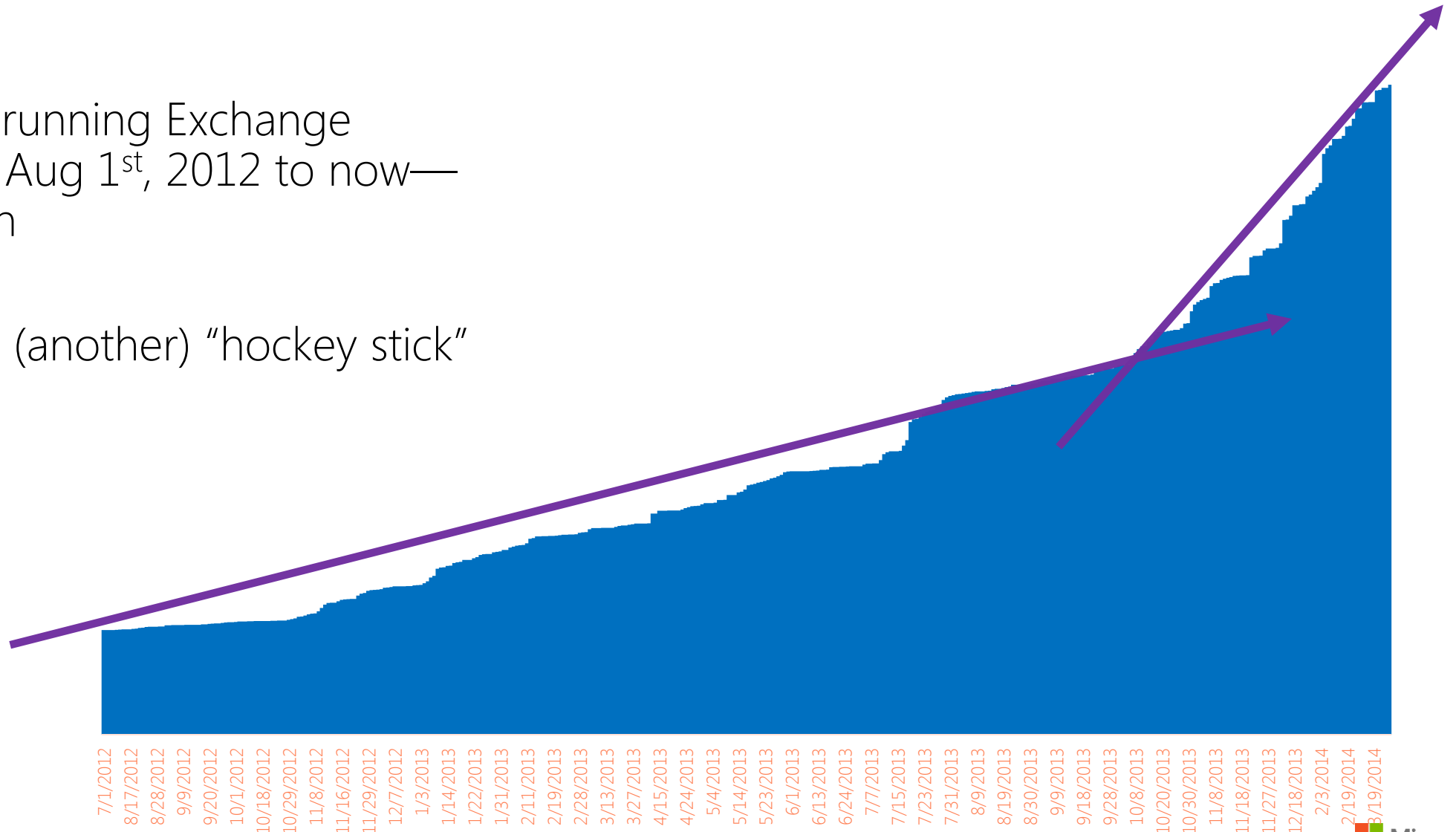
- Proximity to customers
- Energy, Fiber Infrastructure
- Skilled workforce

- 1+ million servers
- 100+ datacenters in 40+ countries

Global Scale: Capacity growth

of servers running Exchange Online from Aug 1st, 2012 to now—
600% growth

Likely hitting (another) “hockey stick”
in growth

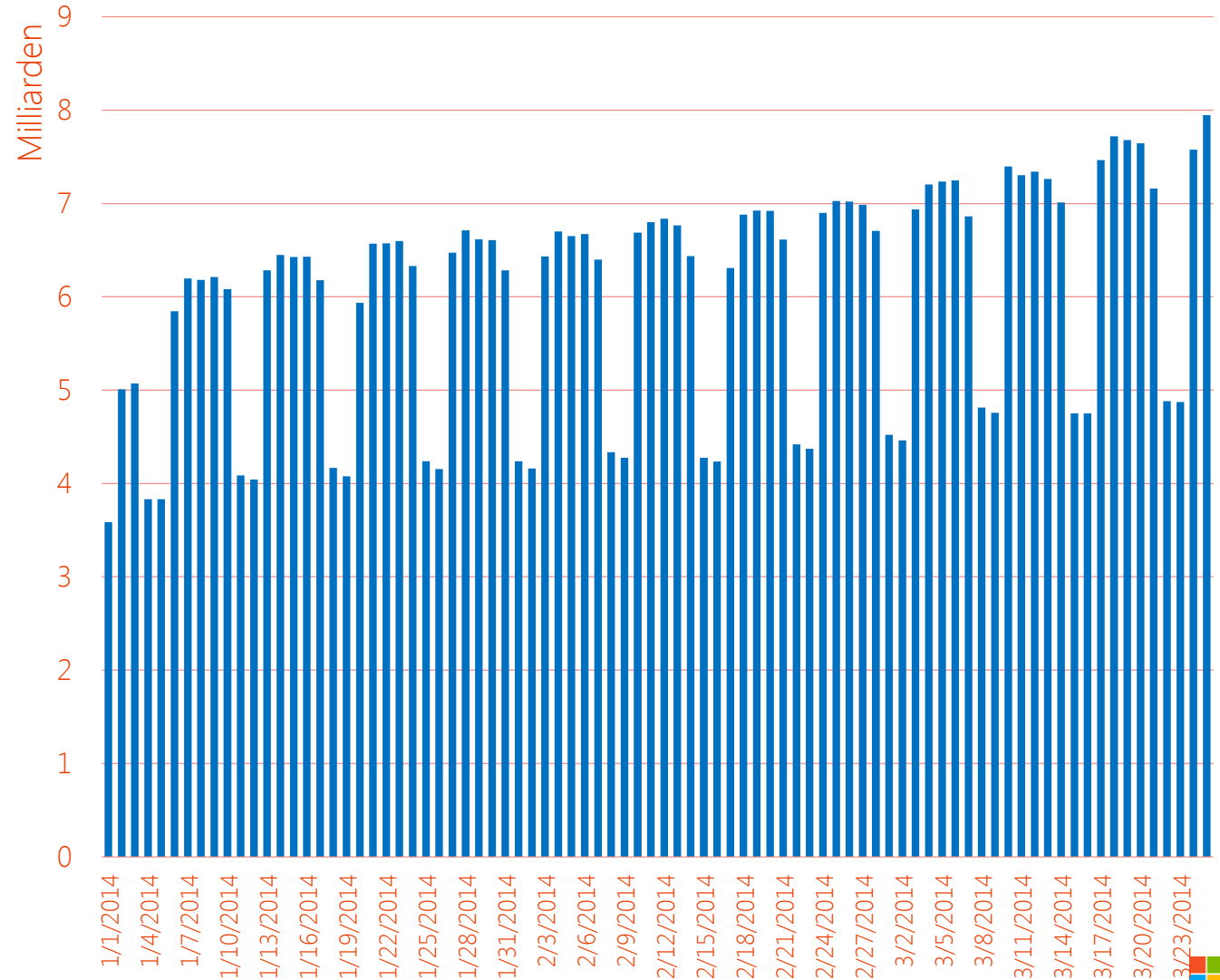


Global Scale: Transaction growth

End user authentication transactions went from 5 billion to 8 billion (62%) in a 3 month period

(Some of it was self-inflicted)

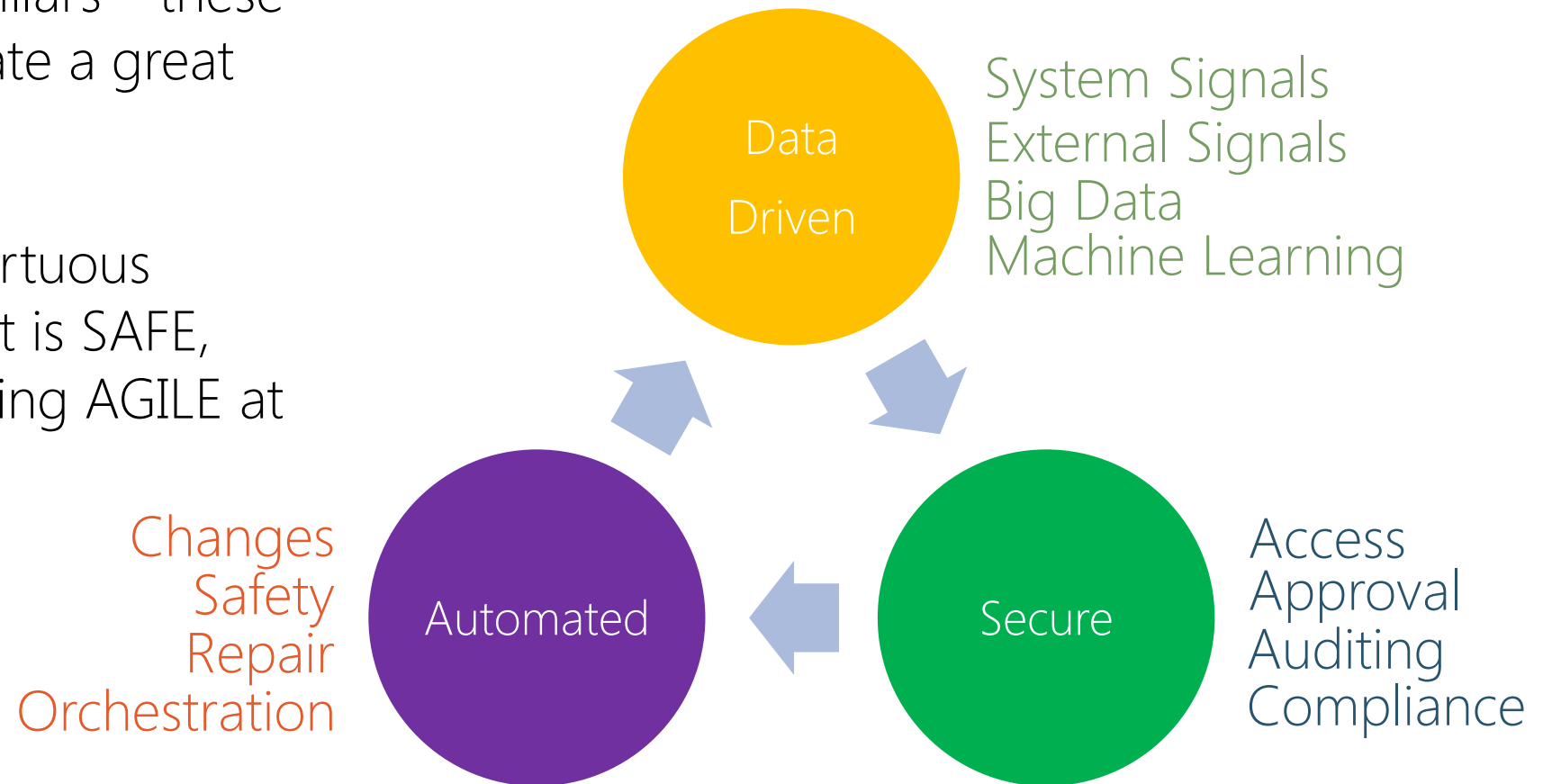
Our product and service infra has to handle spikes and unexpected growth



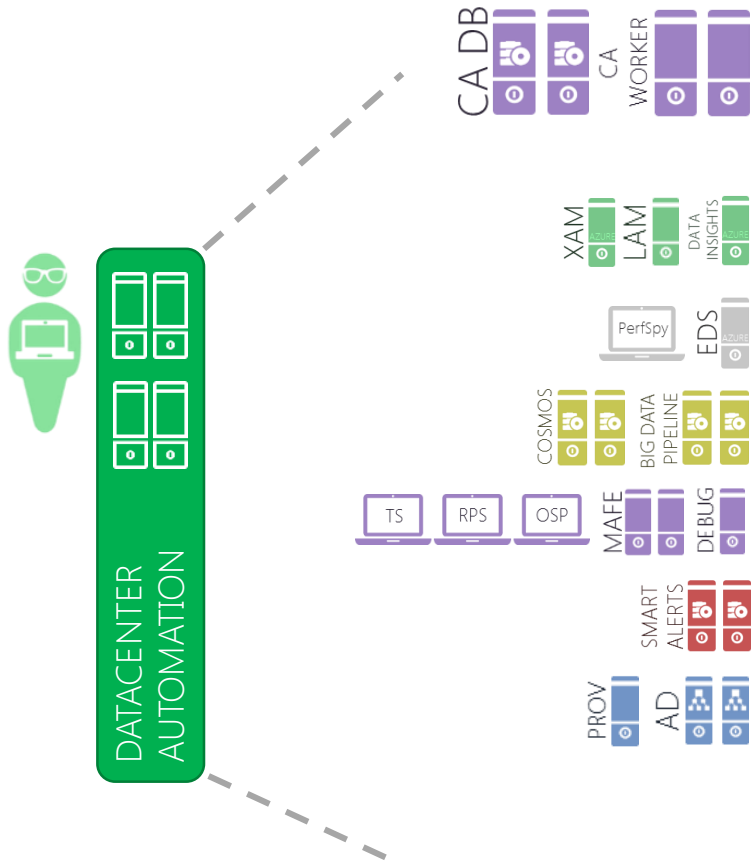
All of it boils down to three pillars

We simplify by focusing all our work along the three pillars—these work in tandem to create a great service fabric

Allows us to create a virtuous automation system that is SAFE, DATA DRIVEN while being AGILE at very high scale



Zoom in: our *Service Fabric*



Orchestration	Central Admin (CA) , the change/task engine for the service
Deployment/Patching	Build, System orchestration (CA) + specialized system and server setup
Monitoring	eXternal Active Monitoring (XAM): outside in probes, Local Active Monitoring (LAM/MA): server probes and recovery, Data Insights (DI): System health assessment/analysis
Diagnostics, Perf	Extensible Diagnostics Service (EDS): perf counters, Watson (per server)
Data (Big, Streaming)	Cosmos , Data Pumpers/Schedulers, Data Insights streaming analysis
On-call Interfaces	Office Service Portal, Remote PowerShell admin access
Notification/Alerting	Smart Alerts (phone, email alerts), on-call scheduling, automated alerts
Provisioning/Directory	Service Account Forest Model (SAFM) via AD and tenant/user addition/updates via Provisioning Pipeline
Networking	Routers, Load Balancers, NATs
New Capacity Pipeline	Fully automated server/device/capacity deployment

is made up of a lot of *stuff*

Data Driven

using signals to improve the service via our “Data Insights Engine”

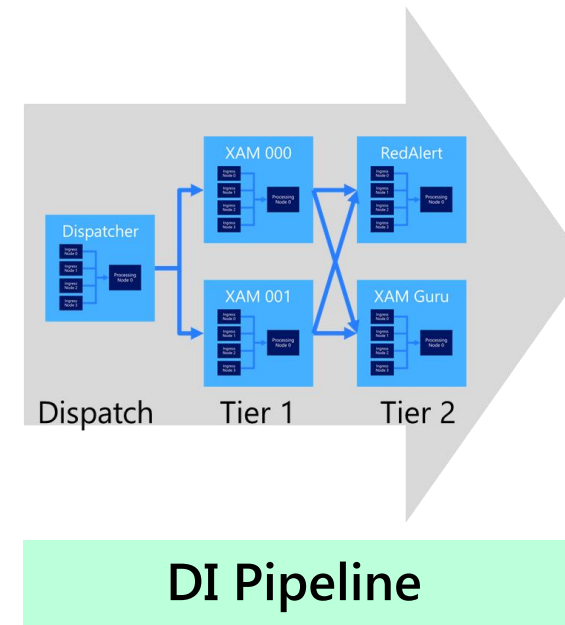
Data Insights Engine

Has to process 100-500 million events/hour—and growing every day

Highly purposed to collect, aggregate and reach conclusions

Built on Microsoft Azure and SQL Azure

Uses latest streaming tech similar to storm, spark



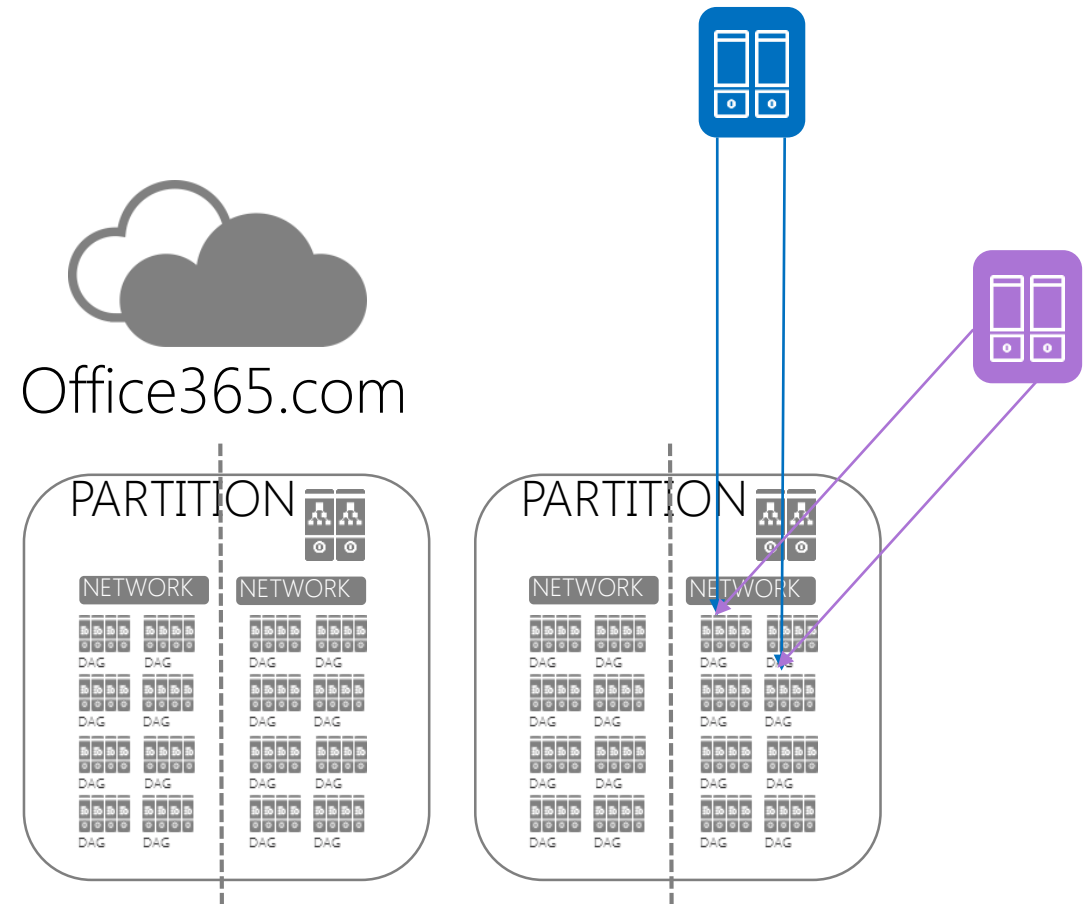
Signals: Outside-In monitoring

Each scenario tests each DB WW
~5mins—ensuring near continuous
verification of availability

From two+ locations to ensure
accuracy and redundancy in system

250 million test transactions per
day to verify the service

Synthetics create a robust “baseline”
or heartbeat for the service



Signals: Usage based

Aggregated error signals from real service usage

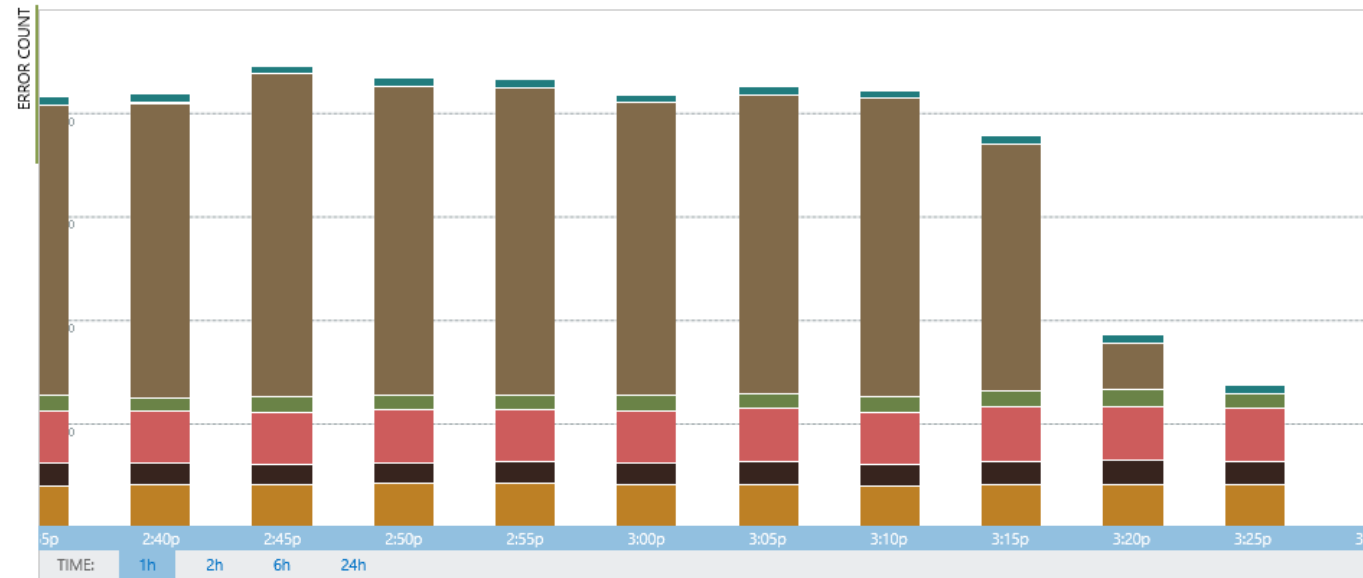
Tells us when something is wrong at the user/client level

Allows us to catch failure scenarios that we didn't anticipate

region: all forest: all dag: all

basic auth

Updated: 03/10/2014 03:31 PM (PST)



SEARCH

Last 15 mins data

ERROR NAME	SERVER	ERROR COUNT
LiveServerUnreachable	BN1PR09CA008	3,396
LiveServerUnreachable	BLUPR06CA036	3,179
LiveServerUnreachable	BY2PR07CA035	2,670
LiveServerUnreachable	BY2PR06CA059	2,108
LiveServerUnreachable	BL2PR09CA002	2,032
LiveServerUnreachable	BLUPR06CA034	1,863
LiveServerUnreachable	BLUPR08CA019	1,745
LiveServerUnreachable	BN1PR02CA004	1,743
LiveServerUnreachable	BLUPR09CA002	1,545

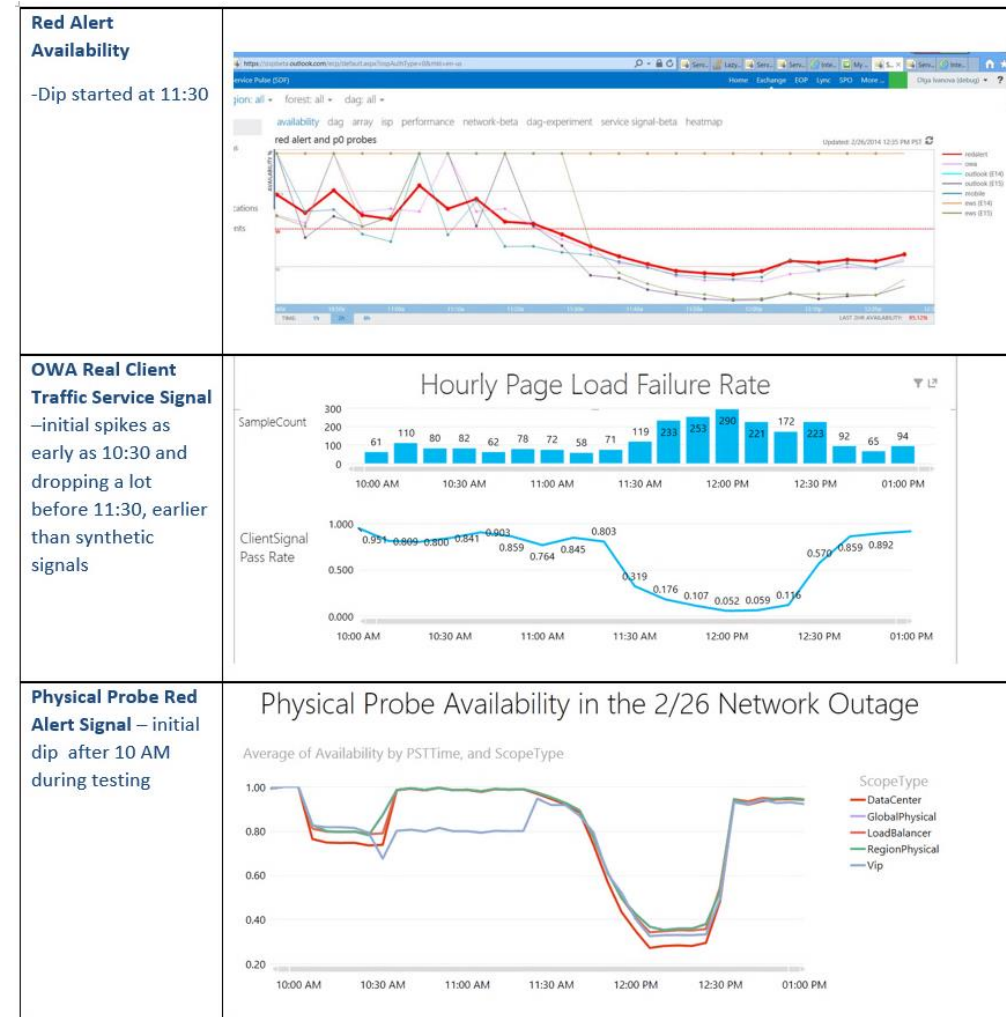
Build confidence via multiple signals

If something is happening across many entities/signals, then it must be true

Apply “baseline” from outside-in as a source of truth

If each signal has reasonability fidelity—you get **~100% accuracy**

We use this technique to build “**Red Alerts**”




Combining these signals lets us triangulate

Combining signals, errors and scopes we can tell what is wrong and where

Dramatically reduces MTTR as no one has to fumble around for root cause

Allows automated recovery—with high confidence comes incredible power

Sent: Wednesday, March 12, 2014 3:42 PM
To: Sanjay Mahida; Yi Wang
Cc: Exchange Data Insights - SDF Alerts
Subject: Smart Alert: RED ALERT: System Level Issue Detected in NAMPDT01.SDF.EXCHANGELABS.COM Forest

 OFFICE SERVICE ALERT attention required

potential impact	service scope
Tenants 0	Forest nampdt01.sdf.exchangelabs.com
Active Users 0	
Total Users 0	
Top Tenants	

description
SDF Health Index - 0.91

A system level issue has been detected because the following failure has been detected:

Multiple Component Failure

The following components have reached a critical level in the past 15 mins:

- E15 Owa Probe availability is 65% (Success Count: 23 ; Failure Count 12) ([OSP link](#))
- E15 EWSGenericDagE15 Probe availability is 100% (Success Count: 35 ; Failure Count 0) ([OSP link](#))
- E15 ActiveSync Probe availability is 100% (Success Count: 35 ; Failure Count 0) ([OSP link](#))
- E15 OutlookCtpE15 Probe availability is 100% (Success Count: 33 ; Failure Count 0) ([OSP link](#))

description

A failure has been detected in EWS component due to the following error increase by 55.8096670221969%: ErrorMailboxStoreUnavailable (Microsoft.Exchange.Data.Storage.MailboxOfflineException)

- EWS ([OSP link](#))

The following servers are associated with the highest number of errors:

- Server: DM2PR03MB509 (100% of all failures)

Auto-posting to health dashboard

4:46 PM is when the alert was raised

Allows us to inform customers in real-time

Keeps engineers focused on recovery

Improves transparency with support and others who keep customers happy

nam	[Throttled] RED ALERT: System Level Issue Detected in NAMSDF01DG030 Dag	12/06 04:46 PM	Data Insights
nam	[Throttled] EAM-EXD-001D-WC: Outlook Health Set unhealthy (OutlookCtp GeneralLowGrade Monit...	12/06 04:46 PM	MOMT, DOMT, XSO
nam	[Throttled] RED ALERT: System Level Issue Detected in NAMSRO1DG051 Dag	12/06 04:46 PM	Data Insights
nam	[Resolved] RED ALERT: System Level Issue Detected in NAMSRO1DG050 Dag	12/06 04:46 PM	Data Insights-ashpre
nam	EAM-EXD-001D-EC: Outlook Health Set unhealthy (OutlookCtp GeneralLowGrade Monitor/MSF1/CH...	12/06 04:45 PM	MOMT, DOMT, XSO-abah...
nam	[Scheduled] [AD health set unhealthy (HealthManagerWorkItemQuarantineMonitor) - [Workitem "Liv...	12/06 04:42 PM	Directory and Liveld Auth-...
nam	[Single Agent Pending] EAM-EXD-001D-WC: Outlook Health Set unhealthy (OutlookCtp GeneralLow...	12/06 04:40 PM	MOMT, DOMT, XSO-Pendi...
mgmt	[Scheduled] DataInsights health set unhealthy (NodeRecycledMonitor/XSI-EXO-PRE-SYS) - Machines ...	12/06 04:35 PM	Data Insights-AshPre
msf	[Resolved] RED ALERT: System Level Issue Detected in MSFT Forest	12/06 04:31 PM	Data Insights-ashpre
nam	[Single Agent Suppressed] EAM-EXD-001D-WC: Outlook Health Set unhealthy (OutlookCtp GeneralL...	12/06 04:22 PM	MOMT, DOMT, XSO-Pendi...
nam	[Scheduled] Search health set unhealthy (ProcessProcessorTimeWarning.noderunner#indexnode1/no...	12/06 04:20 PM	Search-michwils
nam	[Scheduled] [AD health set unhealthy (HealthManagerWorkItemQuarantineMonitor) - [Workitem "Liv...	12/06 04:20 PM	Directory and Liveld Auth-...

EX3497

Status: Investigating

Severity: Sev0

Start Time (UTC): Saturday, December 7, 2013 12:46:34 AM

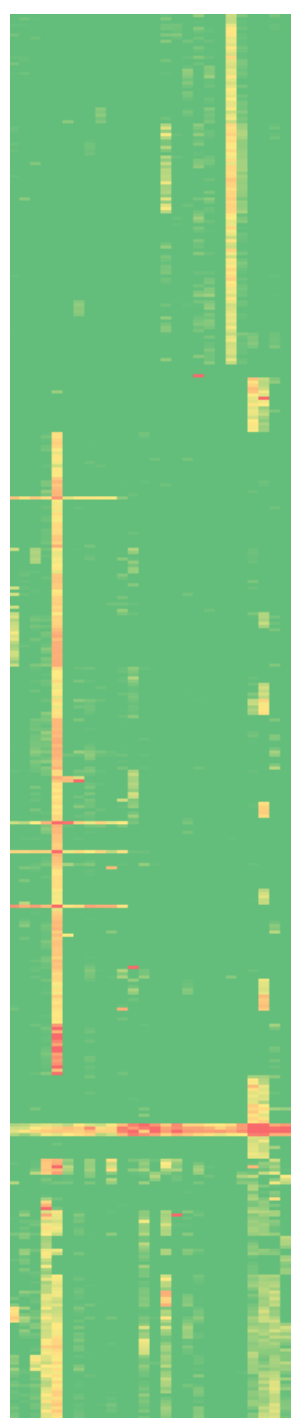
End Time (UTC):

Feature Name: E-Mail and calendar access

Incident Location: namsdf01

Communication:

12/7/2013 12:46:48 AM (UTC) We are investigating a service alert. At this time we don't have enough information to identify whether this is an actual service incident. We will provide more information shortly.



Action

doing work @ massive scale via "Central Admin" (CA)

All actions through "CA"

Why? Changes can be easily destructive, specially if done directly by humans

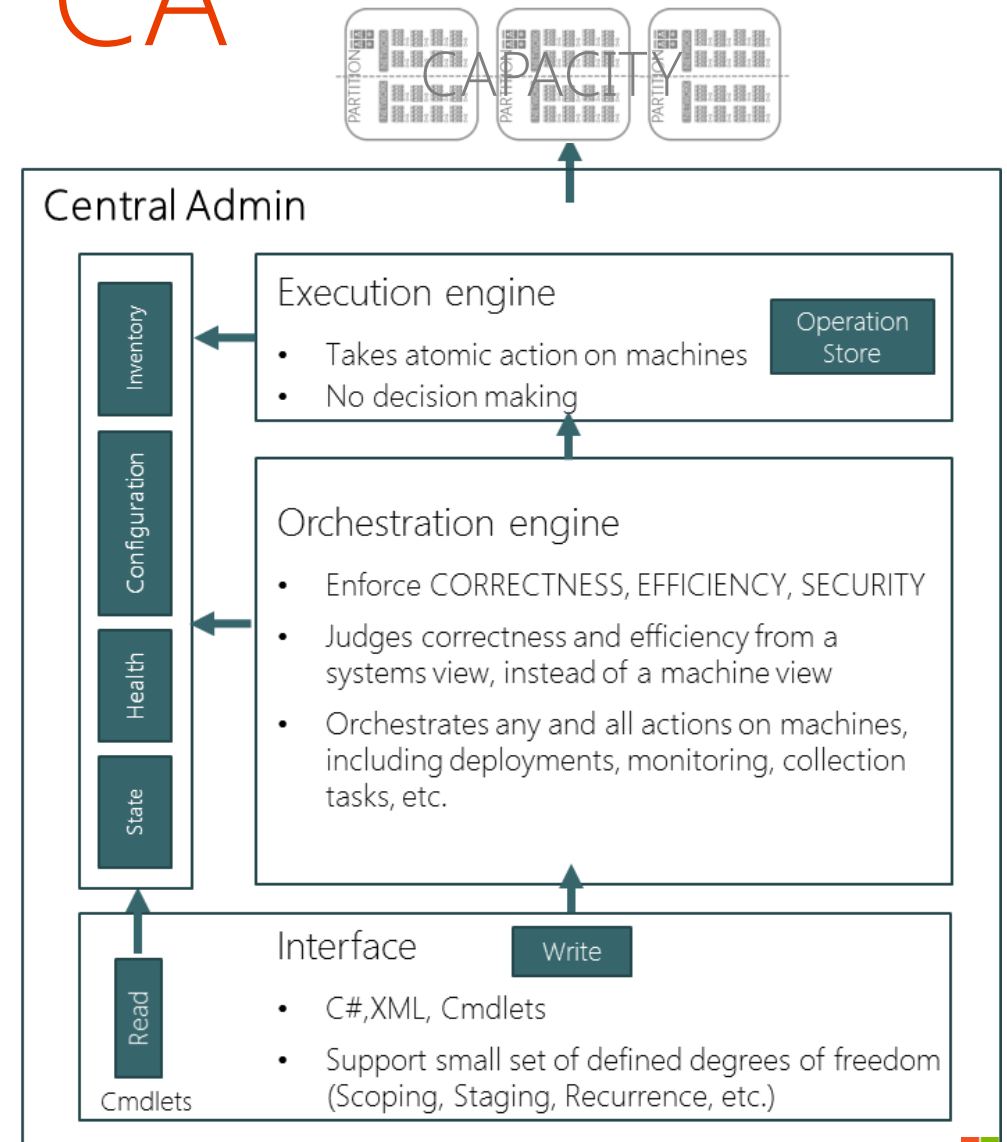
Solution: conduct all actions via a safe, reliable, high throughput system

Central Admin is essentially the "brain" operating the service

Engineers express intent to CA via code (C# "workflows") or PowerShell

Engine then safely applies change across the desired scopes

Data from DI Pipeline informs CA to ensure safety

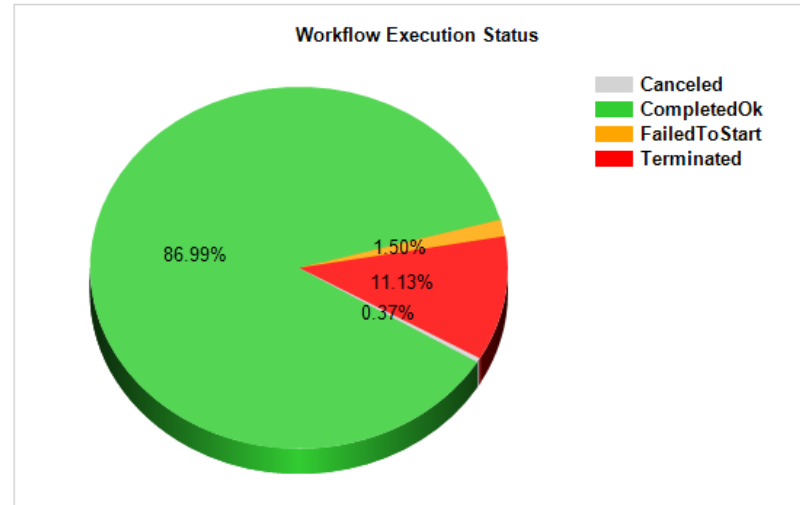


CA at work: workflows for service mgmt.

All actions are built using 'workflows', e.g. deployment, repair, recovery, patching

Even higher order work is done in CA, e.g. rebalance a DAG automatically

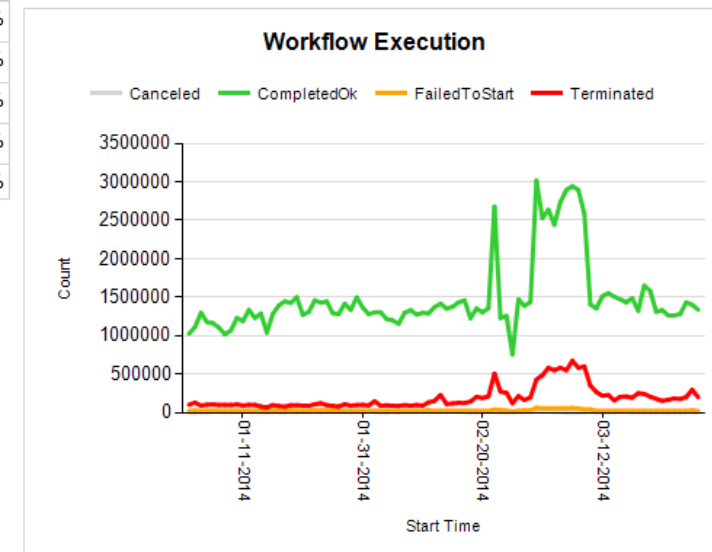
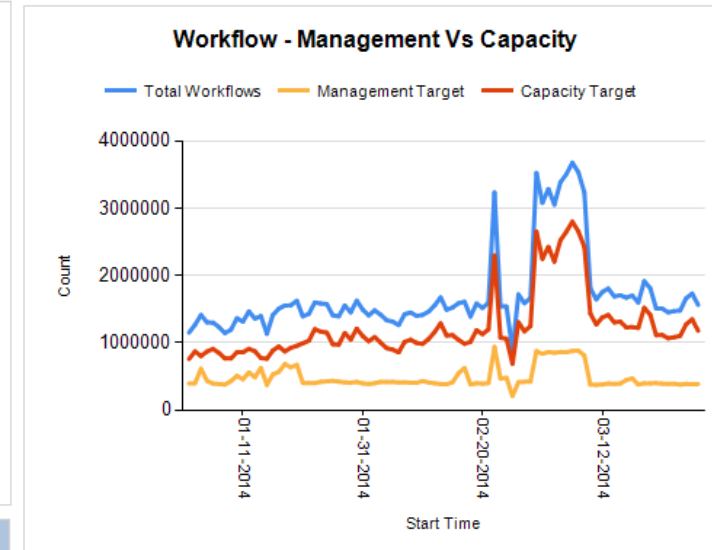
Every month we run **~50 million workflows**. The system is robust enough to handle failures without human intervention



OPERATION STATE	COUNT	%
CompletedOk	127738517	86.99%
FailedToStart	2209107	1.50%
Canceled	550440	0.37%
Terminated	16340364	11.13%
Total	146838428	100.00%

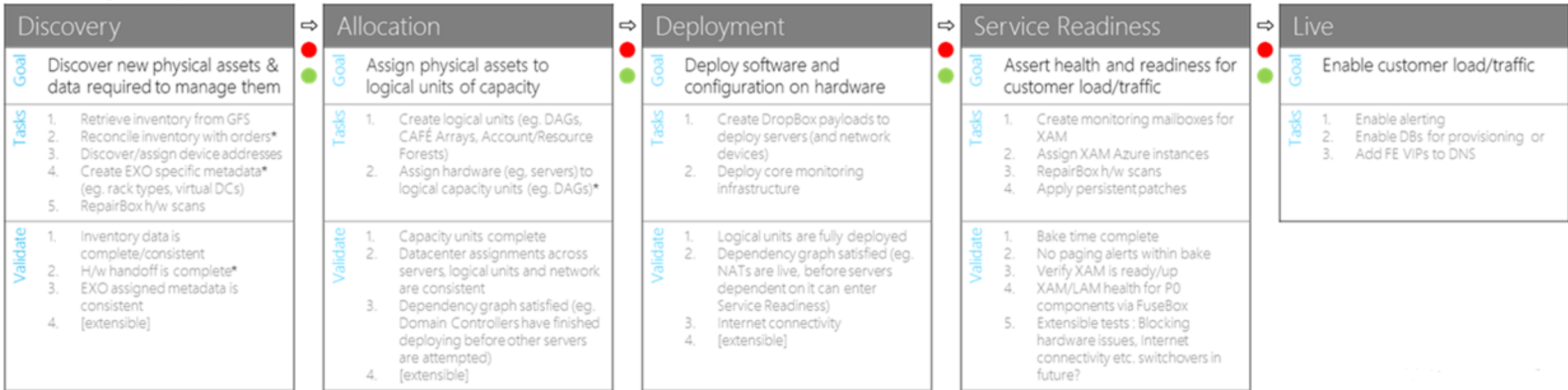
Workflow Distribution:

Management Target	Capacity Target	Total Workflows
41,287,545	105,550,883	146,838,428



New Capacity Pipeline

assembly line phases



Using the systems approach, we shrunk **months to DAYS** to add new capacity

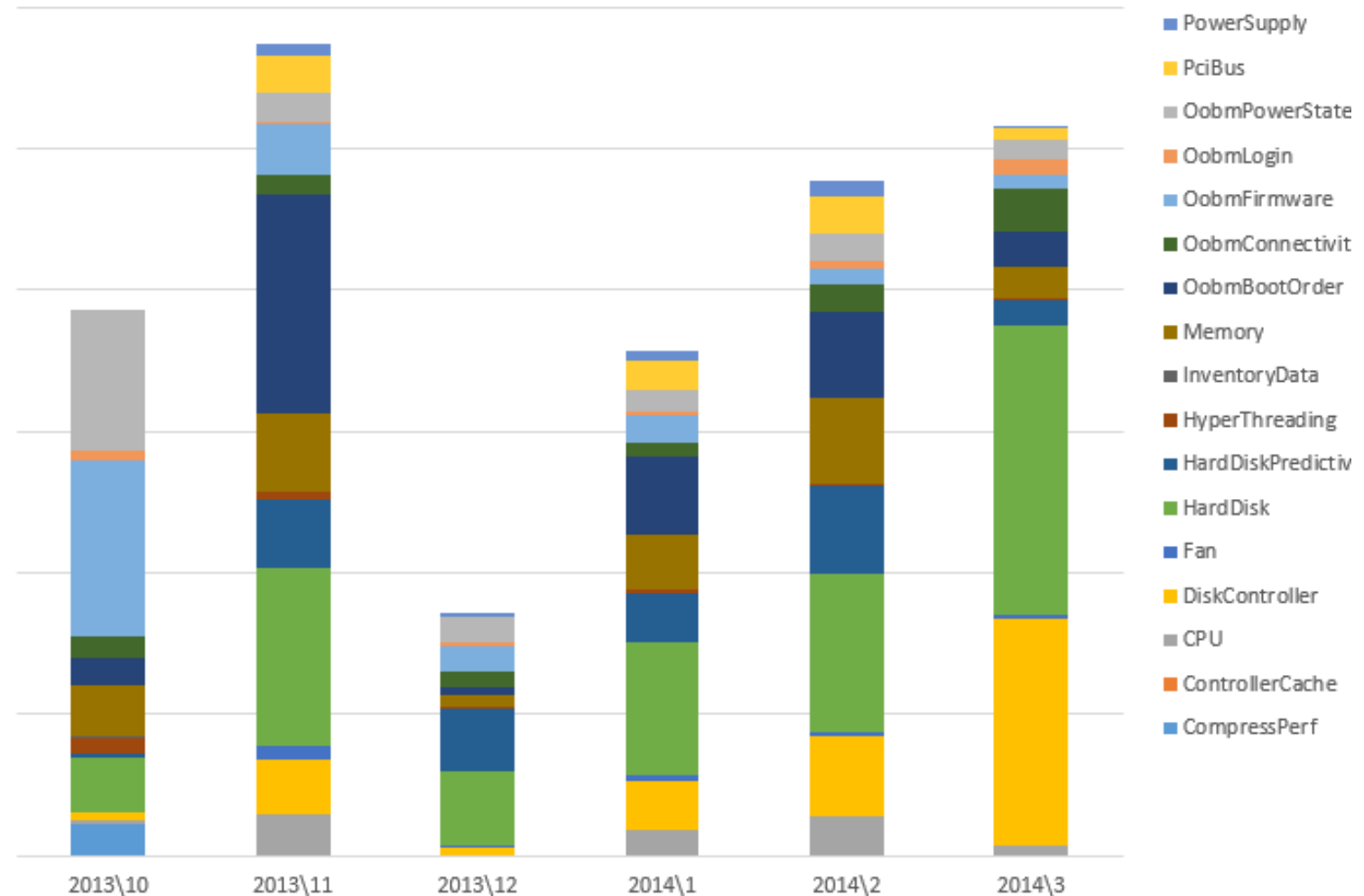
Pipeline built on CA and rich Data systems, including DI for “service green-lighting”

But even after capacity, work doesn't end

We fix/replace 10K or more HW issues every month

Repair == detect + triage + ticket + track completion + bring back to service

HDD, Controllers are top issues—for good reason, that's where the bulk of the hard work is happening



Speaking of repair... "Repairbox"

Specialized CA WF that scans and fixes variety of service issues

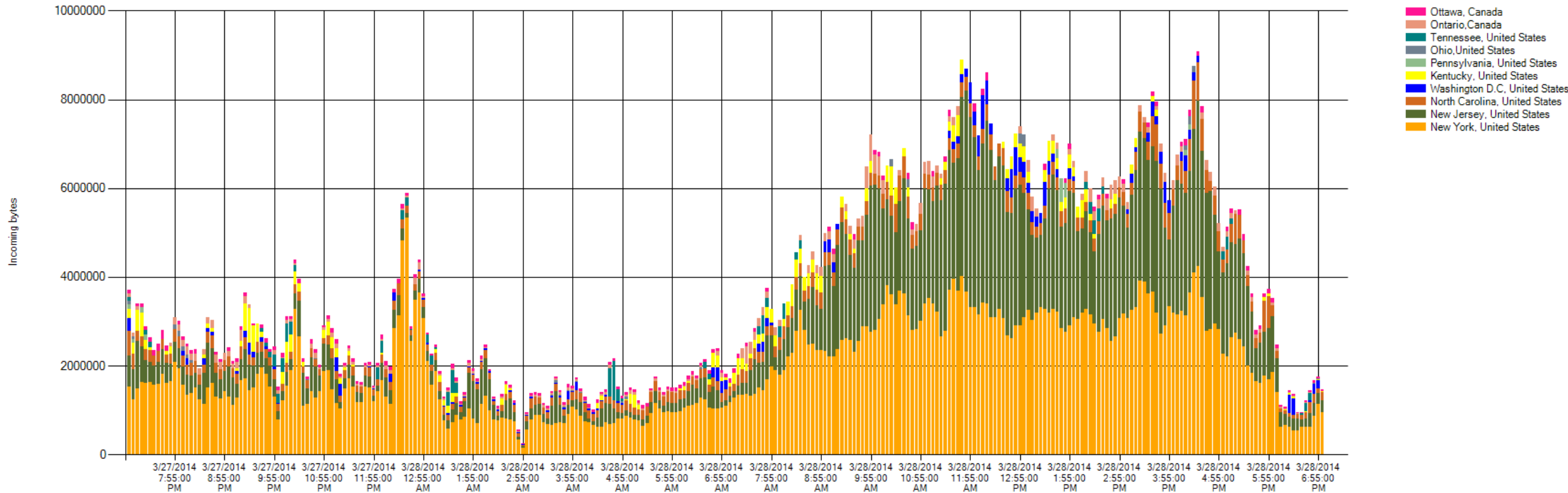
- Consistency checks (e.g. member of the right server group)
- HW repair (automated detection, ticket opening and closing)
- NW repair (e.g. firewall ACL)
- "Base config" repair such as hyper-threading on/off

Our solution to scale along with number of servers, deal with stragglers etc.

Tickets Opened:	1278
Tickets Closed:	1431
Tickets Currently Active:	196
% Automated Found:	77%
Average time to complete (hrs):	9.43
95 Percentile (hrs):	28.73

	Live Capacity	New Capacity	Total
<i>Issue</i>			
HyperThreading	398	44	442
HardDisk	195	30	225
PPM	105		105
WinrmConnectivity	96	1	97
Memory	53	10	63
HardDiskPredictive	39	14	53
Motherboard	41	2	43
NotHW	34	4	38
DiskController	28	9	37
PowerSupply	16	6	22
CPU	9	13	22
OobmBootOrder	19	2	21
Other	18	3	21
ILO IP	12	4	16
ILO Reset	14	2	16
Fan	10	3	13
NIC	9	2	11
InventoryData	4	2	6
NIC Firmware	5		5
ILO Password	1	4	5
OobmPowerState	5		5
Cache Module	4	1	5
High Temp	2	1	3
PSU	2		2
Cable	1		1
Spare		1	1
Total	1120	158	1278

Network: it's a precious resource



Network failures are the **worst** to troubleshoot/fix

Seen everything from ISP/peering failures, cable cuts (freeze!), network gear failing due to software bugs
Our job is to try and get ahead of failures and/or fix fast via failovers. We automatically failover for NAT or VIP failures—no human involved

Putting it all together

Making engineers responsible and responsive

Our Service Philosophy

Principles:

- Engineering first—processes help but are not the only solution
- High emphasis on MTTR across both automated and manual recovery
- Use of automation to “run” the service wherever possible, by the scenario owners
- Direct Escalation, Directly Responsible Individual as the default model
- Low tolerance for making the same mistake twice
- Low tolerance for “off the shelf” solutions to solve core problems
- Bias towards customer trust, compliance and security
- All of this backed by rigorous, hands on attention to the service by **entire team**
- MSR (Monthly Service Review – scrutiny on all service aspects)
- Weekly IM hand-off (managers)
- Monthly Service Readiness review (track customer satisfaction)
- Component level hand-offs, incident/post-mortem reviews (everyone)

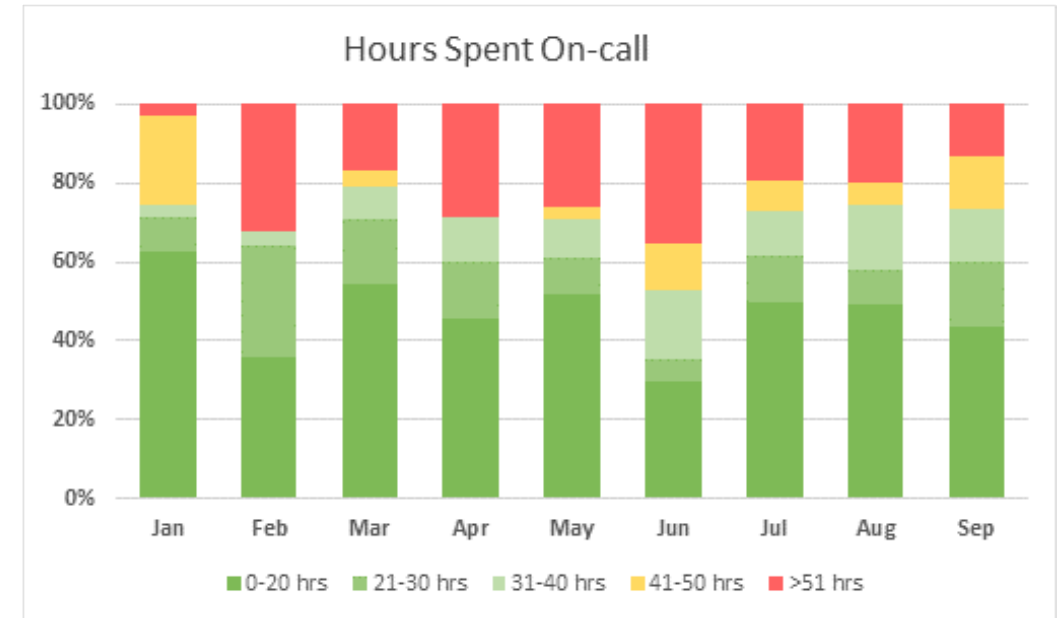
Roles and responsibilities

On-call Engineers: everyone is expected to be on-call across engineering roles

Incident Managers: Team managers are expected to act as senior individuals as needed for incidents

SLT Rotation: "exec IMs" for when customers are impacted

Communications Management: updates comms to customer portal, conduit between support and engineering



September People Impact

1. 176 unique on-calls were paged
2. 33 of them got > 15 pages (40% of pages)
3. 30 got ≥ 8 and ≤ 15 (35%)
4. 113 < 8 pages (15% of pages)

We are here to serve you

The investments we make allow us to continuously improve the service for everyone

We have virtuous cycles in place to learn from any issues, prevent them in the future

Any technology that makes core Exchange better (scale, auto-healing, features) are shipped to on-premises

Core product/on-premises sees more validation at higher stress on an ongoing basis

