



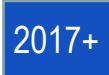
# Removing Silos While Developing A Comprehensive Hybrid Cloud Resiliency Solution

November 08, 2017

Kim A. Eckert

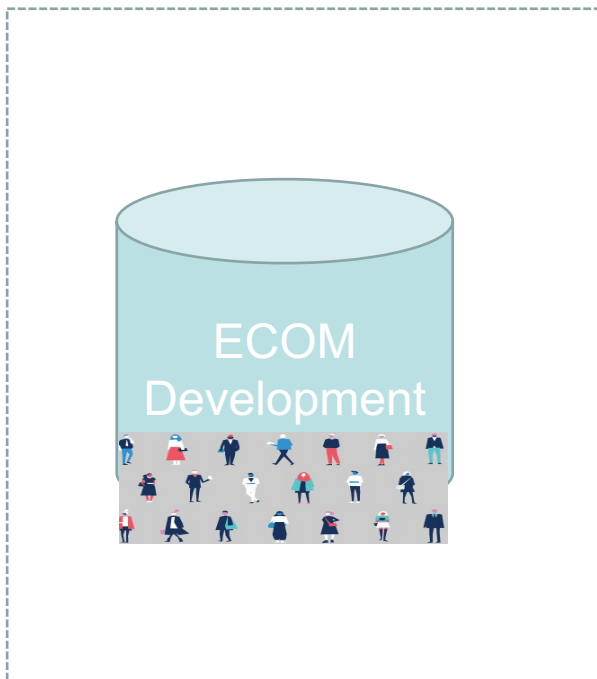
IBM Senior Technical Staff Member; Chief Architect

## Goals

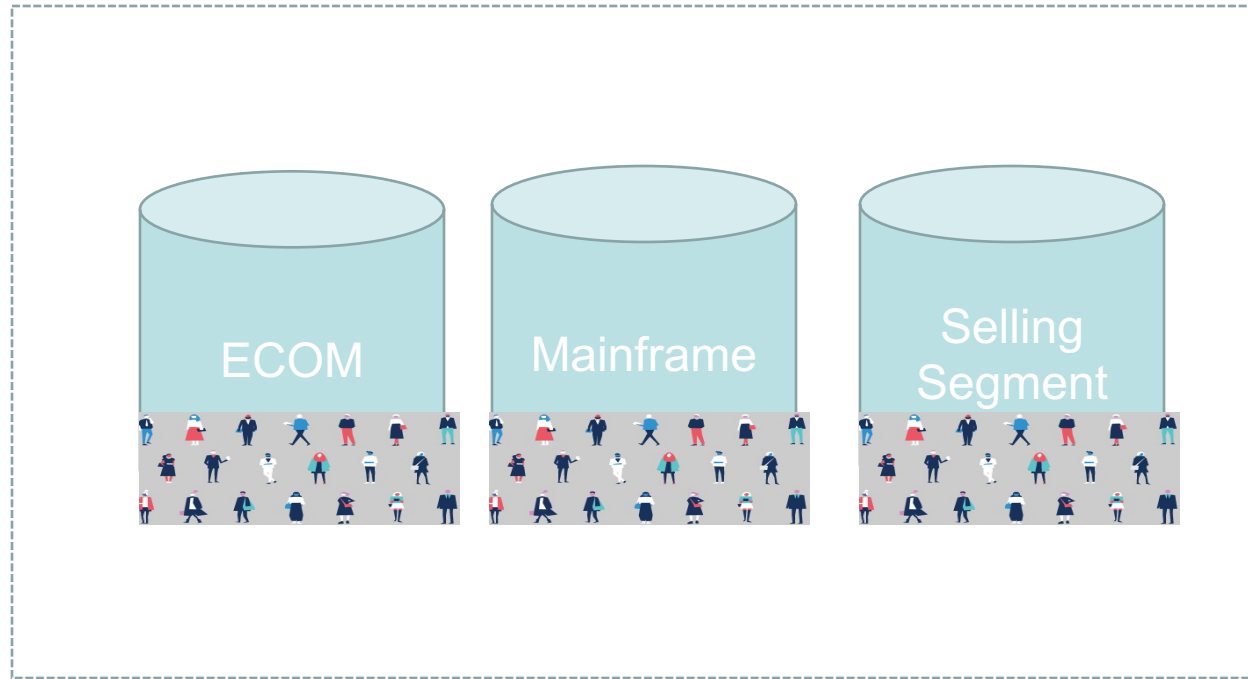
- **Holiday 2017** 
  - ECOM Stability
    - Operational Process recommendations
    - Application/Infrastructure Architecture recommendations
  - Selling Segment (SS) Stability
    - Operational Process recommendations
    - Application/Infrastructure Architecture recommendations
  
- **Path to Public Cloud** 
  - Holiday 2017
  - Post Holiday 2017 
  
- **Being more predictive, machine learning and self-healing**

# Client Organization

## West Coast



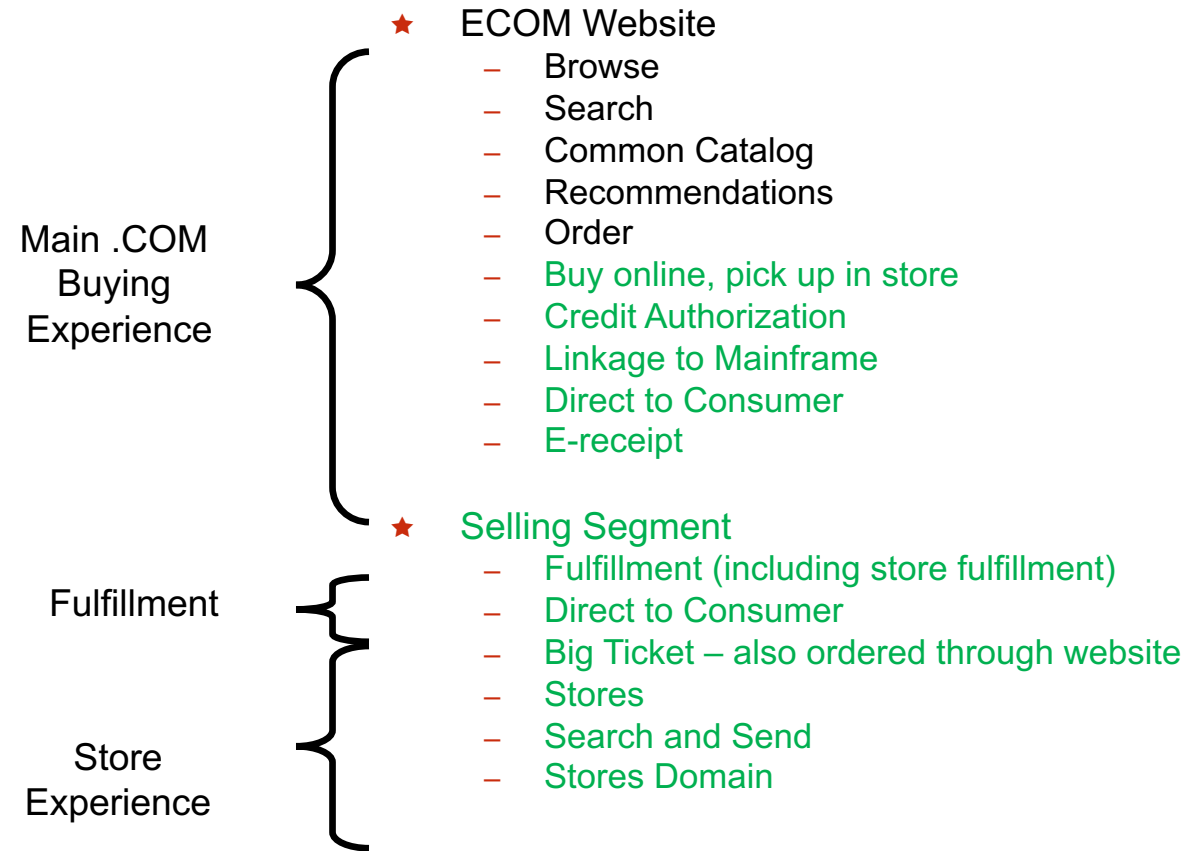
## East Coast



# Scope Reviewed; Applications and Breakdown

There is not clean separation between ECOM and SS.

They are intertwined and require holistic review. (Ecom dependencies on Selling Segment applications are shown in Green.)



Legend

.COM Apps
SS Apps

# Holiday 2017 Stability Recommendations

# Overall Summary of Holiday 2017 Stability Recommendations

- 10 Architectural recommendations
  - 5 critical recommendations
    - VM/Cluster/Storage isolation
    - Network redundancy
    - Additional burst capacity
  
- 16 Operational recommendations
  - 6 critical recommendations
    - Change governance/communication
    - Security/access control
    - End-to-end diagnosis
    - Backups

# Architectural Recommendations Summary of Holiday 2017 Stability

2017

- 10 Architectural recommendations
  - 5 critical recommendations
    - VM/Cluster/Storage isolation
    - Network redundancy
    - Additional burst capacity

# Critical Architectural Risks and Recommendations for Holiday 2017

#	Risk	Mitigation	App
A1	<b>Availability:</b> Current Cell A and Cell B on-prem application domains are sharing common VMWare clusters. Cell A and Cell B sharing storage and/or VMWare cluster fault domains increases the impact of an infrastructure failure.	Create new clusters isolating storage amongst 2 fault domains Place Cell A and Cell B applications on separate fault domains (expand current ESXi redesign) <a href="#">*Additional details on Slide 9.</a>	.COM, SS
A2	<b>Availability:</b> On-prem (DC1) storage under ESXI hosts going down bringing down multiple VM's in diff. clusters. Cell A and Cell B sharing storage and/or VMWare fault domains increases the impact of an infrastructure failure.	Replace with new storage. Treat failure as a site failure and roll over to alternate site until recovery is complete. <a href="#">*Additional details on Slide 9.</a>	.COM, SS
A3	<b>Availability:</b> Lack of physical public cloud NFS-based storage placement for Browse apps.	Work with Cloud team through tickets to validate logical/physical storage mappings	.COM, SS
A4	<b>Availability:</b> Due to static routes and restrictive firewall rules between ECOM and SS domains, alternative network routes cannot be easily established if failures occur on the primary network.	Initiate a discussion w/the various application and network teams to work through application routing changes in case of network failure. <a href="#">*Additional details on Slide 10</a>	.COM, SS
A5	<b>Availability:</b> 2-site model availability during planned and unplanned outages 2-site model (100% capacity) has no room to grow if holiday's are > than expectation	Get Performance/test infra production ready (eg: register monitoring agents, validate NW connectivity) Potentially use Perf/test site as 3 <sup>rd</sup> site to support browse traffic capacity to if one site has issues.	.COM



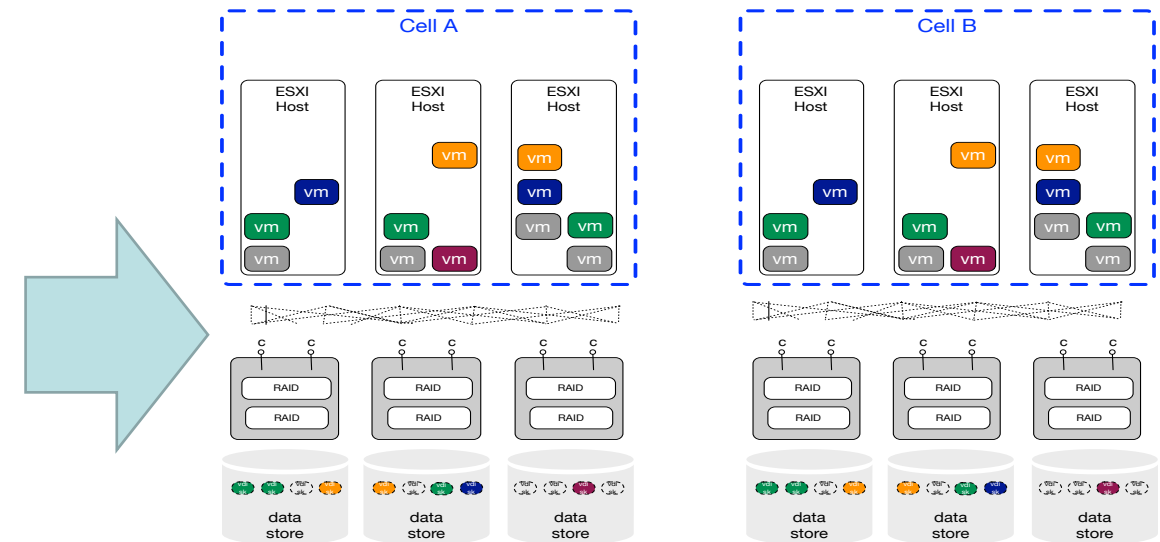
# Additional Architectural Risks and Recommendations for Holiday 2017

#	Risk	Mitigation	App
A6	<b>Availability:</b> ECOM – Cassandra ECOM cluster at risk for BareMetal hardware failures (to reduce RTO for replacement machine).	Order additional bare metal machine w/appropriate config in appropriate data centers (cloud/on-prem)	.COM
A7	<b>Performance:</b> Inconsistent ESXI Storage models (ISCSI/NFS and differing LUN sizes) across Sites introduces variability to assess performance runs.	Performance Test workload on ISCSI vs. NFS (to measure viability of NFS in both locations in future Create clusters in appropriate storage configurations	.COM
A8	<b>Availability:</b> If DC1 fails, SS needs to switch with DC2 (it's DR site), and to connect to DC2 from cloud, SS firewall rules need to be changed.	Initiate a discussion w/the application team to work through application routing changes in case of complete DC1 failover.	.COM, SS
A9	<b>Availability:</b> Cloud to DC1 communication is fragile due to static SS route definitions.	Partner to see what it will take to work through dynamic routing rules from Cloud into SS.	.COM, SS
A10	<b>Problem Diagnosis:</b> Client network has little insight into provider's infrastructure	Get the Client Network team access to the provider's Infrastructure Dashboard	.COM, SS

# Cell deployments are currently a SPOF with infrastructure failures

## Problem:

- Application placement within ESXi infrastructure is not optimized for fault tolerance.
- ESXi review scheduled to work through future placement plan.
- Failure domain extends to storage to host mapping (cell A and cell B) to data store mapping in DC1

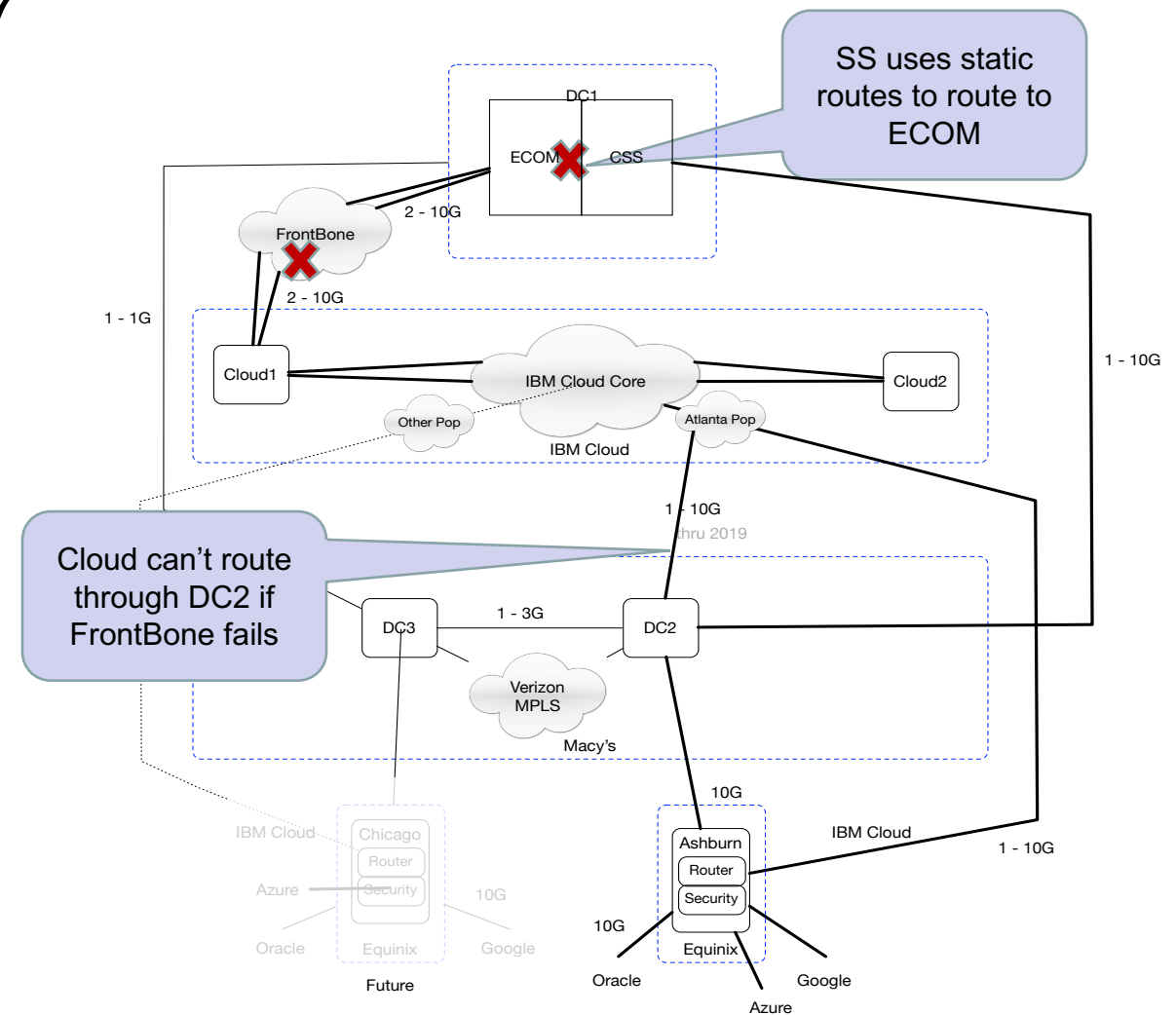


## Recommendation

- ★ Separate fault domains (clusters/storage) per cell
- ★ Spread Apps across clusters and hosts for Availability

# Network Topology Risks for Holiday 2017

- SS uses static routes to get to ECOM/Cloud. If the route fails, alternative routes can't automatically be chosen.
- If FrontBone fails, firewall rules between ECOM and SS domain restricts Cloud routing through DC2 to DC1



# Operational Recommendations Summary of Holiday 2017 Stability

- 16 Operational recommendations
  - 6 critical recommendations
    - Change governance/communication
    - Security/access control
    - End-to-end diagnosis
    - Backups

# Critical Operational Risks and Recommendations for Holiday 2017

#	Risk	Mitigation	App
O1	<b>Problem Determination:</b> Failures for complex applications appear as Network problems. Network ownership is complex	Jointly take inventory of what we can monitor, what we can't, and determine better ways to monitor application flow through the network. Evaluate support options going forward to minimize complexity.	.COM, SS
O2	<b>Communication:</b> Infrastructure changes are not being communicated to all teams. e.g. database team didn't know that servers were being patched	A single Change Review Board will be established in July once Service Now is in place. Vendors have a seat at the table	.COM, SS
O3	<b>Recovery:</b> Cloud DB backups are unreliable as the backups are exceeding backup window	Move backups (Daily and Weekly) to TSM	.COM
O4	<b>Availability:</b> Ensure quick failover for Cassandra DB during Holiday	Will have to approve the project and order necessary equipment (in DC1) - add a bare metal reserve in DC1 and Cloud	.COM, SS
O5	<b>Security:</b> Many people have root access and use shared IDs – this could cause unplanned outages.	Assign a SPOC to manage Client security authorization needs to limit the number of people who have root access to servers. Jointly partner to handle new and existing requests and make appropriate changes.	.COM, SS
O6	<b>Availability:</b> Address automated patch risks.	Change process and automation engine code to improve error avoidance	.COM, SS

# Operational Risks and Recommendations for Holiday 2017

#	Risk	Mitigation	App
O7	<b>Performance:</b> Storage performance in Cloud perf/test ESXi hosts is different than Cloud prod (ISCSI vs. NFS) and adds variability to assessing performance results.	Provide architectural guidance of NFS allocation for endurance/performance storage then have Client assess and implement	.COM
O8	<b>Availability:</b> Unplanned outages are possible due to lack of application failover. E.g. Client had not yet completed an assessment at the time of review.	Joint project started with SS and ECom to assess applications that need failover from DC1 to DC2	.COM, SS
O9	<b>Communication:</b> Including more Development and Architectural teams into the Operational Readiness reviews would improve their effectiveness.	Expand provider's understanding to the architecture/development teams (reduce compartmentalization) Invite infrastructure providers to participate to prepare for infrastructure failures	.COM, SS
O10	<b>Problem Diagnosis:</b> Application dashboards mapped to infrastructure placement do not reflect the latest application deployments in all cases. Potentially slowing or confusing response to critical issues.	Ensure monitoring dashboards for holiday season are up to date on application and host placement. Ensure process to keep this list up to date is implemented.	.COM, SS
O11	New application components getting added are potentially introducing new gaps in planning for the holidays (lack of architectural governance).	Establishing an architectural review board would allow a unified view of application and infrastructure across teams and partners.	.COM, SS

# Operational Risks and Recommendations for Holiday 2017

#	Risk	Mitigation	App
O12	<b>Provisioning:</b> Back and forth build sheet clarification, understanding all the processes followed to ensure a production ready server	Properly fill in newly API-driven submission of server build sheets	.COM, SS
O13	<b>Deployment:</b> Manually managing IP addresses and using duplicate IPs	Implement Infoblox and IPAM for Ecom IP address and DNS management to automate IP and DNS assignment	.COM, SS
O14	<b>Communication:</b> Provider project managers not knowing all the details to backfill each other	Cross-train	.COM, SS
O15	<b>Security:</b> Too many people are accessing VMWare vCenter directly which leads to vCenter risk	Create a jump server and define a process for rolling the users to the jump server.	.COM, SS
O16	<b>Availability:</b> Hardware failures in DC1 causing outages.	Audit of all hardware bios and systems boards	.COM, SS

## Previous Outages and Resulting Recommendations to Enhance Stability

#	Issues (Outages)	Recommendation
I1	Storage Outage	A1, A2,A3
I2	1170 Servers rebooted	O6
I3	Cloud Storage Failure	A2, A3
I4	3 <sup>rd</sup> party provider Network Issue	O1, A8, A9
I5	Cloud Network loss due to changes	O1, A8, A9
I6	Mainframe application outage	O2
I7	Order Drops	O1 <i>(starting point)</i>



# Path to public cloud for Holiday 2017

# Path to public cloud - Recommendations for Holiday 2017

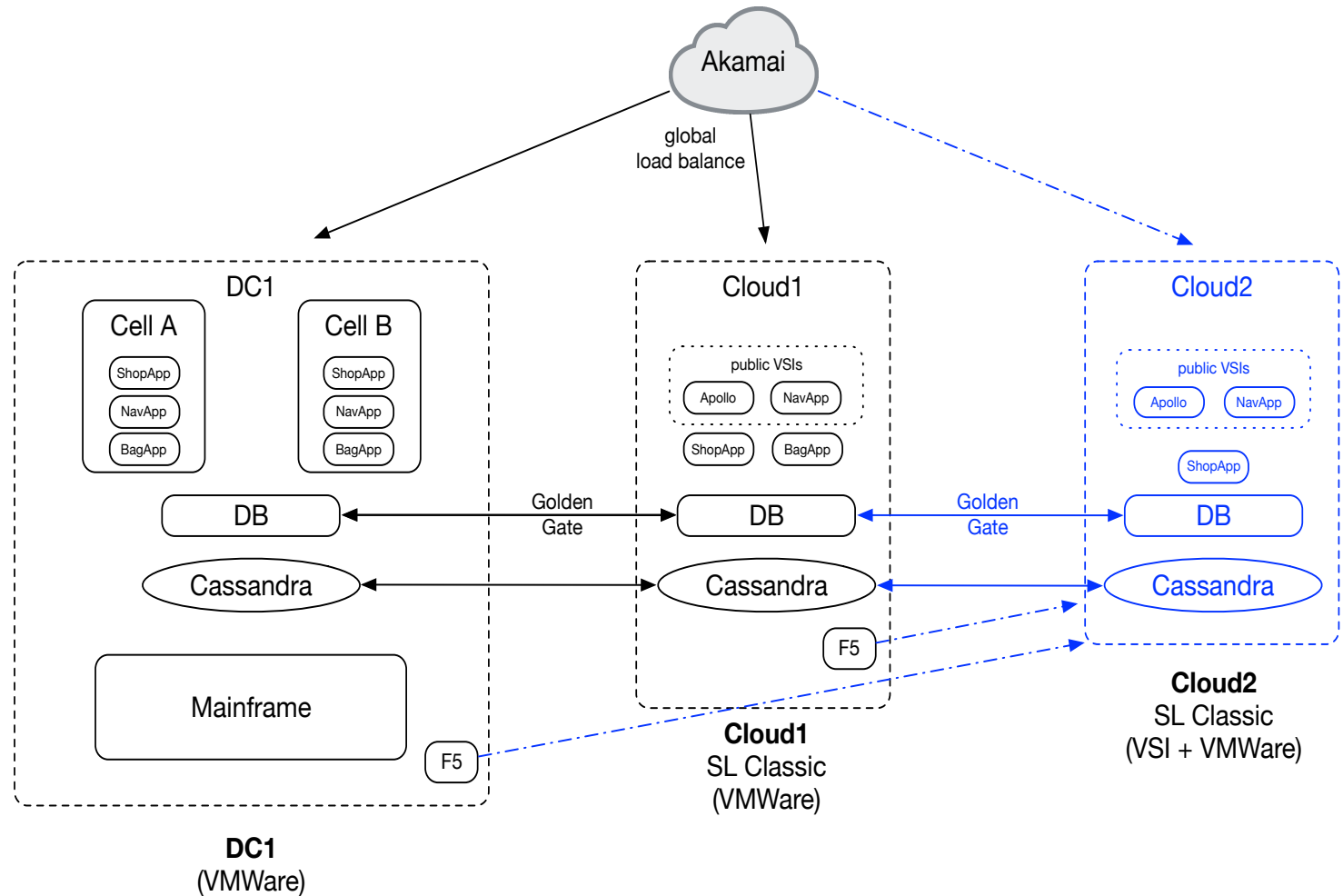
Be able to expand to 2<sup>nd</sup> Cloud instance to handle additional browse traffic

## Benefits/Motivation

- Leverages public VSI capabilities based on Cloud1
- Additional failure domain / availability
- Additional capacity
- Already established based on performance testing
- Allows additional active @ anytime (including maintenance)

## High level changes

- Cassandra sync
- Monitoring agents
- F5 (Akamai) changes



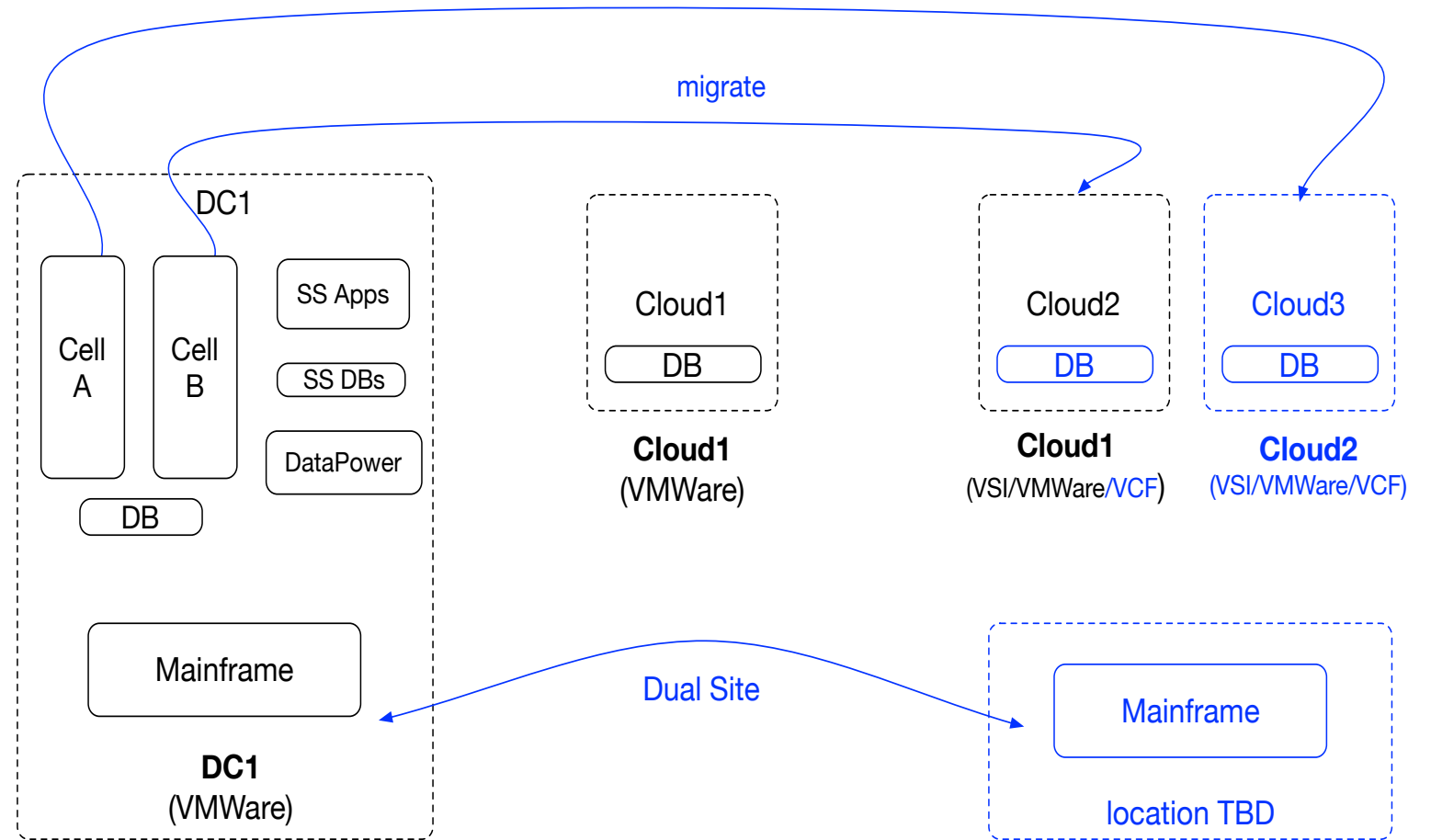
# Path to public cloud - post Holiday 2017

# Architectural Post-Holiday 2017 Recommendations (1h2018)

- Shift from logical cell A/B model to true physical failure domains
- Leverage existing tools/practices, while building new practices/understanding on public cloud.
- Potentially leverage VMWare Cloud Foundation (VCF) to easily stamp out new clusters and simplify migration
- Move away from network appliances to software-based aaS offerings (e.g. security groups, LBaaS)
- Expand to dual zOS site for additional availability

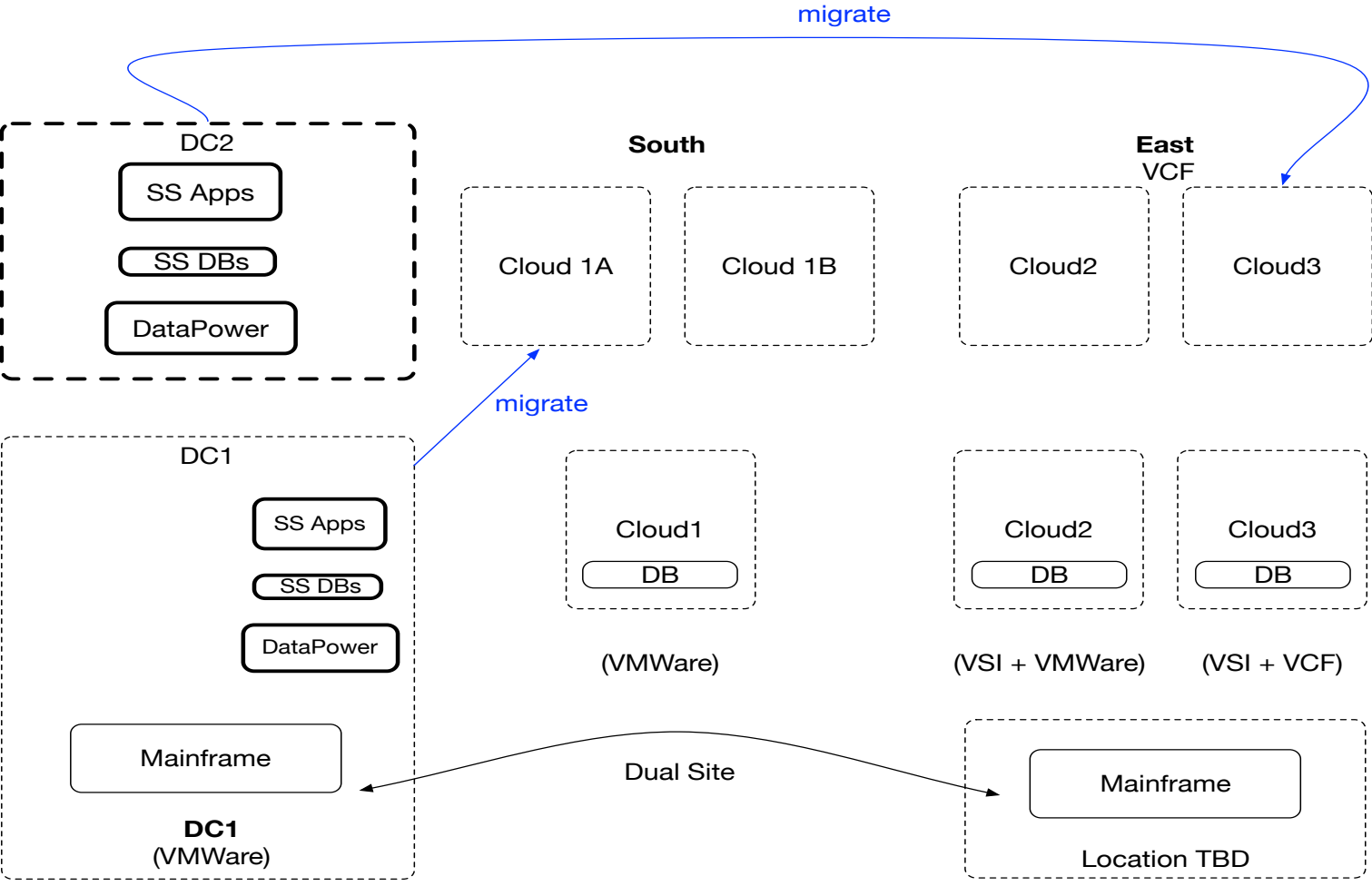
### Benefits/Motivation

- Move away from legacy DC1 to cloud
- 3 datacenter design – (regional Availability across multiple availability zones)
- Full multi-site delivery



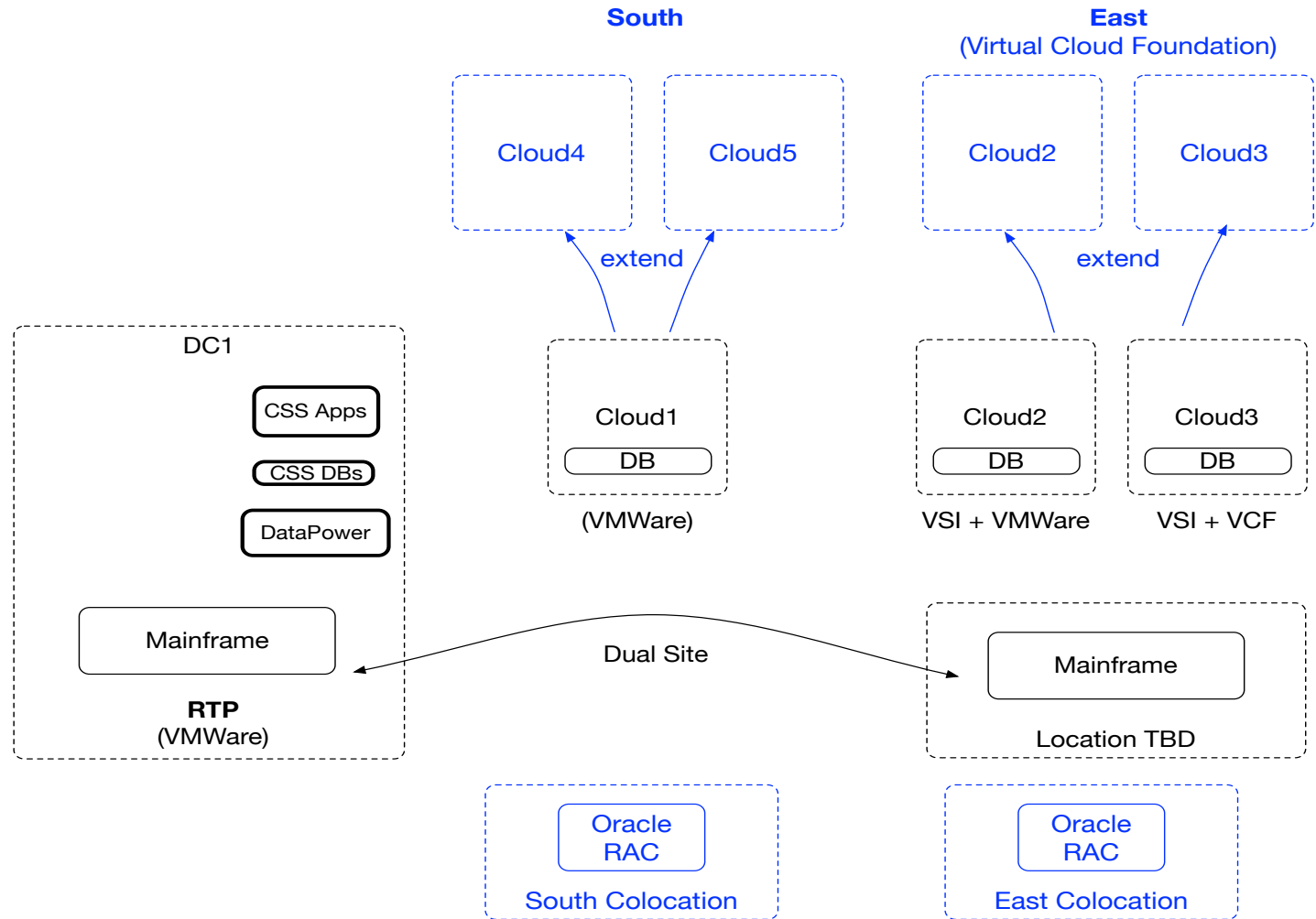
# Architectural Post-Holiday 2017 Recommendations (2h2018)

- 100% cloud native
- Microservices - Docker, Kubernetes
- Expand portions of existing architecture on a as-need basis  
i.e. no forced migration



# SS Shift - Resiliency-based “DC” strategy

- Migrate SS Applications to a cloud-based model
- Shift from a CapEx to an OpEx model
- Lift and shift aspects of cloud vs. rewrite to cloud native
- Host Oracle RAC in Cloud CoLocations



## Additional Observations

- Architects and Developers would benefit from the institution of a more formalized architecture governance model leading to a more complete end-to-end view of how the holistic system operates
- Current Operational Readiness review for Site Operations is a great model to follow to prepare for operational readiness and more stakeholders should engage.
- Most of the fundamentals infrastructure issues seen to date manifest itself as network issues and storage failures that have caused a site failover –
  - Network and Storage would benefit the most from the focus brought by a formal architectural governance model.
- Fault tolerance has been deployed at multiple levels and scopes (but, it seems to come across during the meeting as an art vs. a science).
  - E.g. Multi-site (Akamai)
    - DC1 (F5 and cell), Datapower clusters to failover sites

However, there doesn't seem to be a well-defined policy for when different levels of fault tolerance are invoked.
- Multiple DR approaches and sites make site failover complex
  - E.g. DC1 to DC2, DC1 to DC3, DC1 to DAL has many different load balanced and DR primary to secondary sites.
- There seems to be a number of manual deployment and manual failover across applications
- Cassandra can't take more than a 3-hour outage from one site before things get extremely painful and will have to rebuild the entire Cassandra cluster.

## Exec Summary for Holiday 2017

- Architecturally, the Client is on the right track with multi-site for their application architecture. However, mapping the application to the IT infrastructure model (and how it aligns to a cloud strategy) and eliminating SPOFs (which there still are some) is a work-in-progress
  - E.g. Moving cloud-native, Cassandra..
- Operationally, we have found security, governance, end-to-end monitoring, backup and infrastructure provisioning weaknesses on ECOM
  - E.g., Network diagnosis, change management control, root access to core IT systems
- End-to-End delivery model complexity contributing to outages, problem determination, quick recoverability is extremely complex
  - e.g. lack of end-to-end tracing
- Client and Provider's need to partner to better plan and prioritize improvements for Holiday stability
  - E.g. 3 related projects (ESXi redesign) at varying levels of the stack affecting application placement and isolation. Independently would require 3 change windows vs. 1.



# **A more predictive, machine learning and self-healing infrastructure and operations**

## Predictive Availability

- Leveraging Data Science tools and a team of analysts to review data to develop Predictive availability model and Key Performance Indicators
- Working with over 25 million log entries to generate uniform data relating to system health for supported devices within the environment
- Identify trends in order to predict system behaviors
- Develop preventive measures to respond to predictive signals

# Backup

# Single Point of Failures Identified during Architectural Reviews

2017

- **Critical**
  - **ECOM**
    - **Application - ECOM** – Account xAPI, MSP Client batch, OES-based applications
    - **Infrastructure** - DC1 ESXI Cell clusters have SPOF with storage crashes due to lack of application/infrastructure failure domains
  - **SS**
    - **Application - D2C** – 15 apps are at risk due to DC1 site SPOF
    - **Mainframe**
      - **Mainframe** is a single site (DC1) and is at risk if there are site connectivity failures as well as Mainframe failures
      - **GoldenGate from Mainframe DB2** – single “extractor” from DB2 mainframe
    - **Infrastructure**
      - DC1 ESXI Cell clusters have SPOF with storage crashes due to lack of application/infrastructure failure domains
      - SS Failover for availability of DC1 and/or connectivity from DAL to SS DC1 requires manual work for routing changes and failover

2017+

- **Non-Critical**
  - **ECOM - non-strategic** a few legacy apps are SPOFs.
  - **SS** (*reduced risk due to limited initial implementations*)
    - **Application**
      - **Store fulfillment** – single site in DC2. No HA or DR
      - **SS** – marketing/loyalty – NFS is SPOF
      - **Inventory** (mainframe) is SPOF with DC1 site
      - **Stores transformation** – ECS replication is a single site in DC2
      - **Production Cloud2** - SPOF as no backup site discussed
    - **Infrastructure**
      - **Production Cloud2** - Public cloud storage is local SSD (no ability to recover if disk failures)

## ECOM Architecture

- **SPOFs –**
  - **Critical:** Account xAPI, MSP Client batch, OES-based applications
  - **Non-Critical:** PROs, and a few legacy apps are SPOFs.
- **Detailed Analysis:** 76 ECOM apps
- **Critical:** 56 of 76
- **Monitored:** All but 2 non-critical apps
- **Automated Deployment:** 5 are for QA, Perf, DC1, DAL; 38 are for Perf, DC1, DAL; 17 are for QA, Perf, DAL, 16 are customized or physical servers
- **DR:** All but 2 non-critical apps

**Note:** ECOM is dependent on a number of critical SS Apps

## SS Architecture

- **SPOFs**
  - **Marketing/Loyalty** – NFS is SPOF
  - **D2C** – 15 linux-based apps are at risk due to DC1 site SPOF
  - **IWM** – (Store fulfillment) – single site in DC2. No DR or HA
- **Detailed analysis:** 46 of 375 SS applications
- **Critical:** 9 critical apps of the 46
- **Monitored:** 23 of 46
- **Automated Deployment** – 38 of 46 are manual with one 50% automated
- **DR:** Of the critical apps, MCHUB is the only one that does not have a true DR  
*(this is being worked on now to have DR in DC2 and should be complete before Holiday 2017)*

## Work Effort Completed By The Team

- Held 15 days of workshops with multiple Client teams that produced architectural diagrams and knowledge transfer
- Prompted client to produce architecture artifacts that previously did not exist
- Allowed diverse teams to learn another aspect of the business
- Enabled providers to gain an understanding of client applications and their flows

## Assumptions throughout the document

- All DBs are initially set up in an HA model (e.g. DB2, Oracle, Cassandra)
- All key services w/persistence are set up in HA mode (e.g. Kafka, Tibco, JDG)
- Cross-site ESB (Tibco/Kafka/MQ) replication is not done.  
Instead, all ESB events manifest itself into some datastore (DB2, Oracle, Cassandra) – and replication of those happens cross-site.
- Load testing is representative of a holiday environment (and truly represents the anticipated workload)



## Network Complexities on ownership

- Many of the issues investigated appear as network issues as the symptom is that you cannot reach a dependency
- Problem ownership spans multiple locations, groups, and challenges

### Responsible

- |                |  |
|----------------|--|
| 1. Client      | 1. DC2, Akamai, DC3, DC4   |
| 2. Verizon     | 2. DirectLink / MPLS   |
| 3. IBM         | 3. Cloud NW/VLAN, SLi  |
| 4. ATT         | 4. DC1 Network, Frontbone (CMS), Internet<br>(5+ order drops in DC1) |
| 5. Vital Net   | 5. F5, Vyattas, Gateways   |
| 6. Optiv       | 6. ASM   |
| 7. CenturyLink | 7. Internet  |

- Recommendation:
  - Operational Failure RACI analysis review

## Reviews & Participants

- ECOM Team:
  - (2) Managers
  - (8) Architects
  - (4) DBAs
- SS Team:
  - (4) VPs
  - (8) Directors
  - (6) Managers
  - (4) Enterprise Architects
  - (3) Architects
  - (10) System Specialists
  - (2) Developers
  - (2) Project Managers
- Mainframe Team:
  - (1) Enterprise Architect
  - (2) System Specialists

## DC1 SPOF

- Cell A and Cell B in DC1 are logical artifacts
  - Unclear on whether applications are mapped either in specific clusters (vs. random mappings)
  - Should be in separate “fault domains” of the physical infrastructure (including storage infrastructure)
    - e.g. outage (storage outage in DC1)
    - There have been 2 other SAN controller outages (microcode update problem)
- Benefits
  - Allows fault domains
  - Keeps clarity between Cells (and topologies) across Clouds, DC1

## Burst to Cloud2 for Holiday 2017

- Limit to browse for ECOM
- Performance testing should cover
  - Browsing use cases to help w/the load for Cloud2 (to prepare to see whether or not it can handle the load)
  - Cloud1 will handle 100% of the load (to test that)
  - DC1 should be also validated to do some performance testing to validate F5 changes to offload to Cloud2.
- No GG (or DB replication) for Cloud2