Problem Management Lite
Quarterly Deep Dive

Agenda

➢ **Objectives:**

❑ Start from the Top	10 minutes	**1:05 – 1:15**
❑ High-Profile Problem Review	45 minutes	**1:15 – 2:00**
❑ Metrics	10 minutes	**2:00 – 2:10**
❑ Next Steps	10 minutes	**2:10 – 2:25**

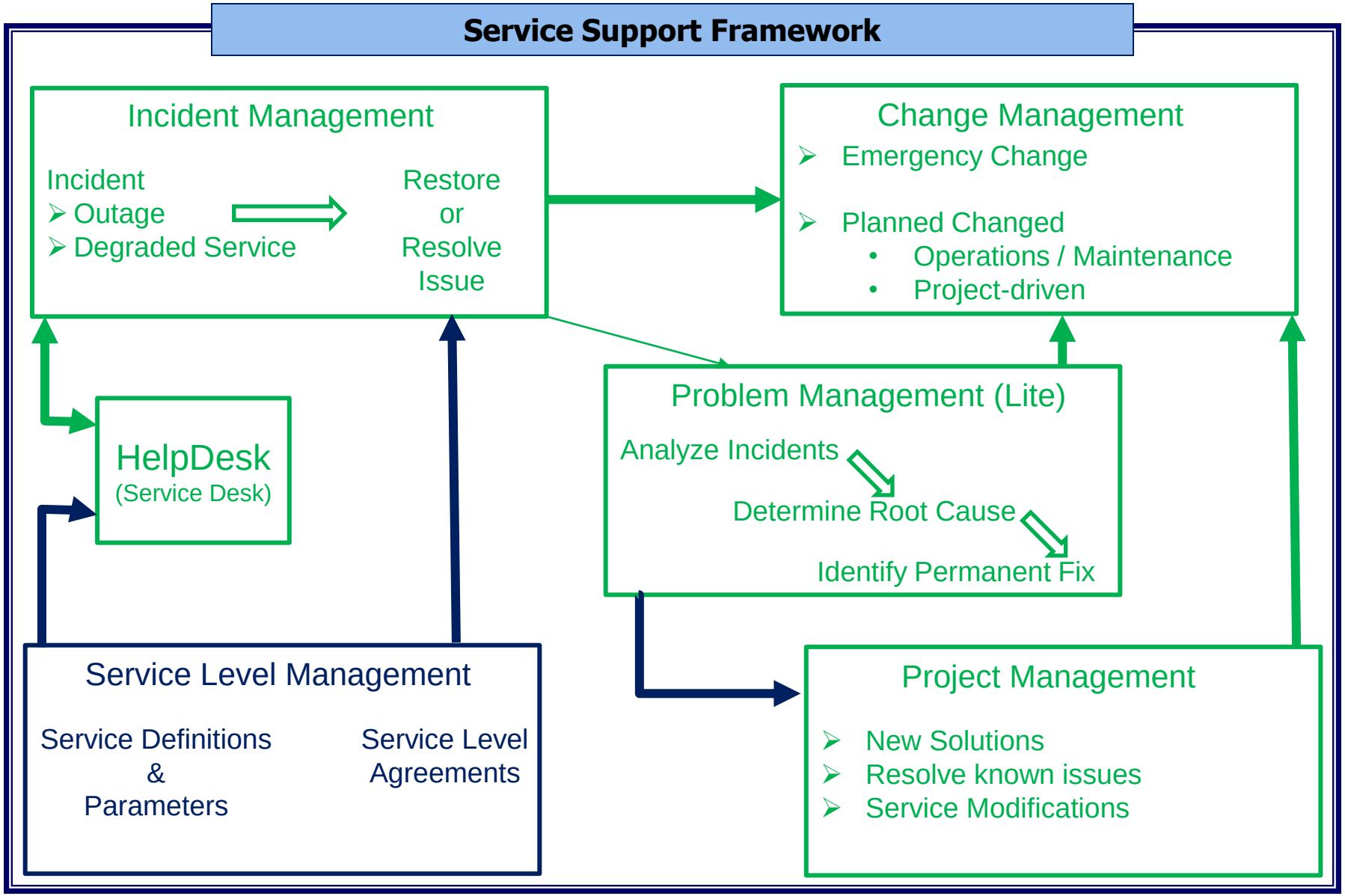
➢ **Goals:**

- ❑ Keep leadership apprised of technical risks
- ❑ Identify next steps

March 28, 2013

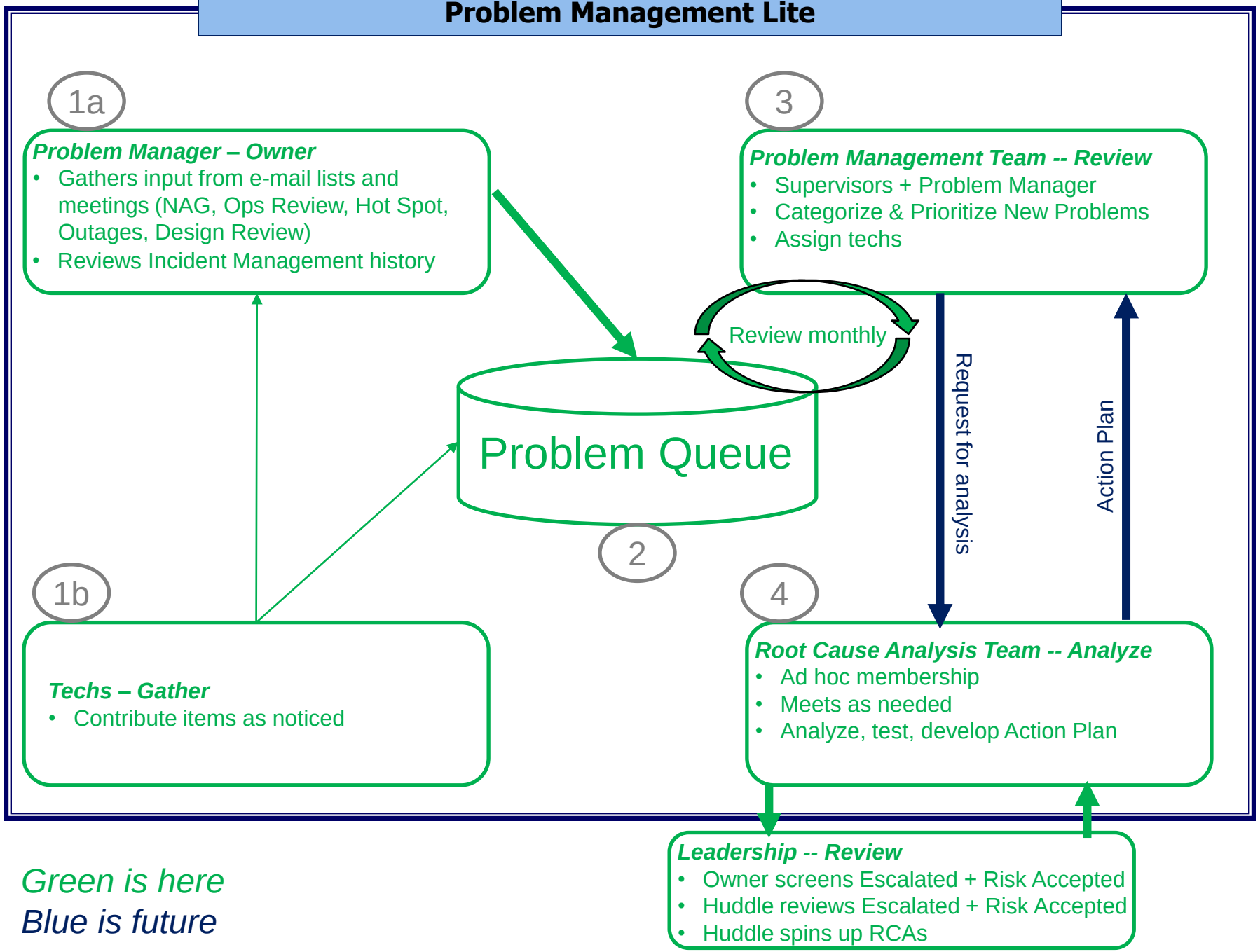
Stuart Kendrick

Problem A cause, or potential cause, of an incident that has already, or may in the future, interfere with a defined IT service



Green is here
Blue is future

Problem Management Lite



Green is here

Blue is future

Successes

In 2012, we Resolved 32 Problems from a pool of ~140

Priority	Title	Time in Queue
2	Both *Silo* NICs plugged into same Ethernet switch	1 month
2	Bug in Firewall OS Will Cause Internet Isolation	2 months
2	Bugs in *Cobalt* OS Could Cause Data Loss	6 months
3	Exchange < = > Zimbra Daylight Savings Time Mismatch	8 months
3	Failing card in Data Center switch	9 months
3	Intermittently rebooting phones in Thomas Building	12 months
3	Security Vulnerability in TSM Backup Client (~100 servers)	18 months
…		

During Q1 2013, we Resolved 10 Problems from a pool of ~115

Priority	Title	Time in Queue
3	CPU Soft Lockup on *Lamprey* (HPC)	3 months
3	Security Vulnerability in Telecommuter Software (hundreds of clients)	3 months
3	Security Vulnerabilities in MMZ Transport Gear (Internet access)	11 months
…		

- *We are Resolving Problems … although typical turn-around time is ~9 months*
- *Every single one of the Resolved Problems has stayed Resolved*

Surprise

#465 Cabrini Pathology

Years of repeated complaints, but none of us thought to create a Problem Record

- Outlook stalls for second to minutes
- Applications become unresponsive
- Browser windows fail to draw
- Reports take tens of minutes to load
- Print jobs take tens of minutes to complete
- DTS unable to Remote Desktop to Cabrini stations

Wolfe Maykut brought this to our attention; we have spun up an RCA, in progress

Future mitigation: Problem Manager reviews Remedy logs monthly, looks for patterns

Transition

Next we delve into the P1 – P2 review

Questions regarding *Start from the Top*?

Snapshot on March 28, 2013

	Open	Assigned	Total
P1	1	1	1
P2	9	0	9
P3	28	5	33
P4	29	4	33
P5	16	7	23
	83	16	99

See All High-Profile Problems

#	Priority	Title	Action
Reviewed during December 2012 Deep Dive			
126	2	Tungsten Stumbles	Stall until OS Upgrade
404	2	Tungsten Performance Degraded	Stall until OS Upgrade
~~406~~	~~2~~	~~SQL Server Crumps During Storage Events~~	Examine during Q1 2013 Harvard/Princeton Migration
24	2	Data Centers Isolated During Router Failover	Test a Sample; Report Back
187	2	Uncertain NetApp Response to Router Failover	Stall until OS upgrade
101	2	Yale Building Routers Intermittently Unresponsive	Point Stuart at this in April
Review Today			
471	1	Domain Controllers Freeze	
455	2	vColo Stumbles During Router/Storage Failover	
457	2	Intermittent Data Protection Gaps	
460	2	Intermittent Failed Network Connections for MIS gear	
460	3	ACD Not Highly-Available	

{Problem Title}

{Mega-Problem Icon}

Priority 1-5	**Status** Open | Assigned	**Owner** CIT Dept	**Service Impacted** Service Catalogue item	**Customer Impacted** End-Users Affected	**Start** Date
Incident Likelihood	High | Medium | Low			{How likely is this to occur?}	
Incident Impact	Widespread | Significant | Limited | Local			{Description of effect on end-users}	
Incident Recovery	High | Medium | Low		{How much effort in terms of staff hours & $$ to recover from an Incident?}		
Root Cause	Known | Unknown			{Description}	

Summary

{Description}

Status - Degrading | Static | Improving

{Is the Problem worsening, staying the same, or getting better?}

{Analysis | Mitigation | Resolution} Plan	**{Analysis | Mitigation | Resolution} Effort -** High | Medium | Low
{What might we do to analyze, mitigate, and/or fix this Problem?}	{Cost in terms of staff hours & $$}

- ❑ Do you understand the risk? If not, what do you need to know?
- ❑ Next steps

Domain Controllers Freeze

Priority	Status	Owner	Service Impacted	Customer Impacted	Start
1	Assigned	InfraOps	Infrastructure Services	FHCRC + SCCA	March 2013

Incident Likelihood	High — Three Incidents across one week: 3/13 mid-day, 3/15 mid-day, 3/17 mid-day
Incident Impact	High — Some applications hang (Outlook, Hyperion, possibly others)
Incident Recovery	Medium — Reboot domain controllers, then application-specific remediation
Root Cause	Unknown

Summary

- Our four Microsoft Active Directory domain controllers provide authentication and directory services
- On three occasions, two of them have frozen simultaneously: the Windows GUI no longer responsive from the console
- In theory, applications should fail over to the surviving domain controllers, making this event transparent to the end-user
- In practice, several applications stumble and become unusable until the boxes are rebooted

Status - Unknown

- We don't have enough history to declare whether this issue is Improving, Degrading, or remaining Static

Analysis Plan	Analysis Effort - Low
Ignore – hasn't recurred for several weeks or Spin up an RCA: install data capture tools	40 hours 2 staff

- ❑ Do you understand the risk? If not, what do you need to know?
- ❑ Next steps: Leave Open | RCA | Risk Accept …

vColo Stumbles During Storage or Network Failovers

Priority	Status	Owner	Service Impacted	Customer Impacted	Start
2	Open	InfraOps	Hosting	FHCRC	June 2010

Incident Likelihood	High
Incident Impact	Widespread — vColo disrupted, various applications unavailable
Incident Recovery	Low — ~2 - 40 CIT + NAG hours across one day, gradual restoration of end-user services
Root Cause	Unknown

Summary

- During routine failover of highly-available Storage or Network services, roughly half of vColo becomes isolated for minutes to hours
- Ditto during unplanned failovers
- Not clear if this is a bug or a limitation in VMware's highly-available capabilities

Status - Static

Analysis Plan	Analysis Effort – Low
Spin up RCA: understand why, sketch options	60 hours 3 staff

- ❑ Do you understand the risk? If not, what do you need to know?
- ❑ Next steps: Leave Open | Risk Accept …

Intermittent Data Protection Gaps

Priority	Status	Owner	Service Impacted	Customer Impacted	Start
3	Open	Architecture	Data Protection	FHCRC	March 2010

Incident Likelihood	Low — Two Incidents in recent memory: June 2011 Fred and November 2011 Silo
Incident Impact	High — Loss of data and/or servers
Incident Recovery	High — ~~50-500 CIT + NAG hours across days to months, gradual restoration of end-user services
Root Cause	Known — Insufficient resources in data protection systems / lack of archiving

Summary

- Our data protection systems no longer backup all our data during nightly incrementals and weekly fulls
- Depending on the timing of a failure, we might lose one day of data (best case), multiple days (likely), or even weeks (worst case)
- The gaps vary day-from-day and by system (vColo, Tungsten …)
- Backups fail intermittently (not clear why); an automated recovery feature allows them to complete

Status - Degrading

- As data volume grows and systems become slower, our gaps become wider
- Long-term resolution will involve a substantial re-engineering effort

Analysis Plan	Analysis Effort - Low
Spin up an RCA to quantify the gaps	30 hours 4 staff

- ❑ Do you understand the risk? If not, what do you need to know?
- ❑ Next steps: Leave Open | RCA | Risk Accept …

Intermittent Failed Network Connections for MIS gear

Priority	Status	Owner	Service Impacted	Customer Impacted	Start
2	Open	InfraOps	Connectivity	FHCRC + SCCA	March 2010

Incident Likelihood	High
Incident Impact	Significant — InfrastructureTasks interrupted; Possible data loss and application corruption
Incident Recovery	Low - High — ~2-200 CIT + NAG hours across one week, gradual restoration of end-user services
Root Cause	Unknown

Summary

- Hosts inside J4-401 are briefly unable to talk with one another
- This leads to failed Infrastructure Tasks, e.g. Backups, Database Maintenance …. and occasional Server Freezes
- Hard to tell how frequent the issue is …
  Infrastructure Tasks tend to be Patient (retry many times) and Automated (no human involved)
- Documented cases occur rarely (~once/month)

Status - Degrading

Analysis Plan	Analysis Effort - High
Spin up RCA	200 hours 4 staff

- ❑ Do you understand the risk? If not, what do you need to know?
- ❑ Next steps: Leave Open | RCA | Risk Accept …

ACD not Highly Available

Priority	Status	Owner	Service Impacted	Customer Impacted	Start
3	Open	InfraOps	Voice	FHCRC	2007

Incident Likelihood	Low
Incident Impact	Medium — 'Fast Busy' for these numbers for 2 hours – 1 week
Incident Recovery	Low — ~1 hour – 1 day of CIT staff time
Root Cause	Known — Design Choice

Summary

- We provide an Automated Call Distribution (ACD) service for two numbers
  667-5000 (FHCRC's main number) and 667-5700 (CIT's HelpDesk)
- An ACD service allows multiple handsets to ring simultaneously when a call arrives, allowing multiple people to service the number
- We have only one server (*Audhumla)* behind this capability
- When *Audhumla* fails (hasn't happened) or is isolated (happens occasionally), calls to x5000 and x5700 receive 'Fast Busy'

Status - Degrading

- The entire voice environment is degrading: we are multiple major versions behind; we haven't patched in years (security or stability); the hardware is approaching end-of-life
- I am using Problem Management to verify that this audience knows and accepts this risk

Resolution Plan	Resolution Effort - Low
Stall to see if the FY2014 Voice Refresh proposal gets funded	$xx capital 1-2 staff / yy hours

- ❑ Do you understand the risk? If not, what do you need to know?
- ❑ Next steps: Leave Open | Project | Risk Accept …

Transition

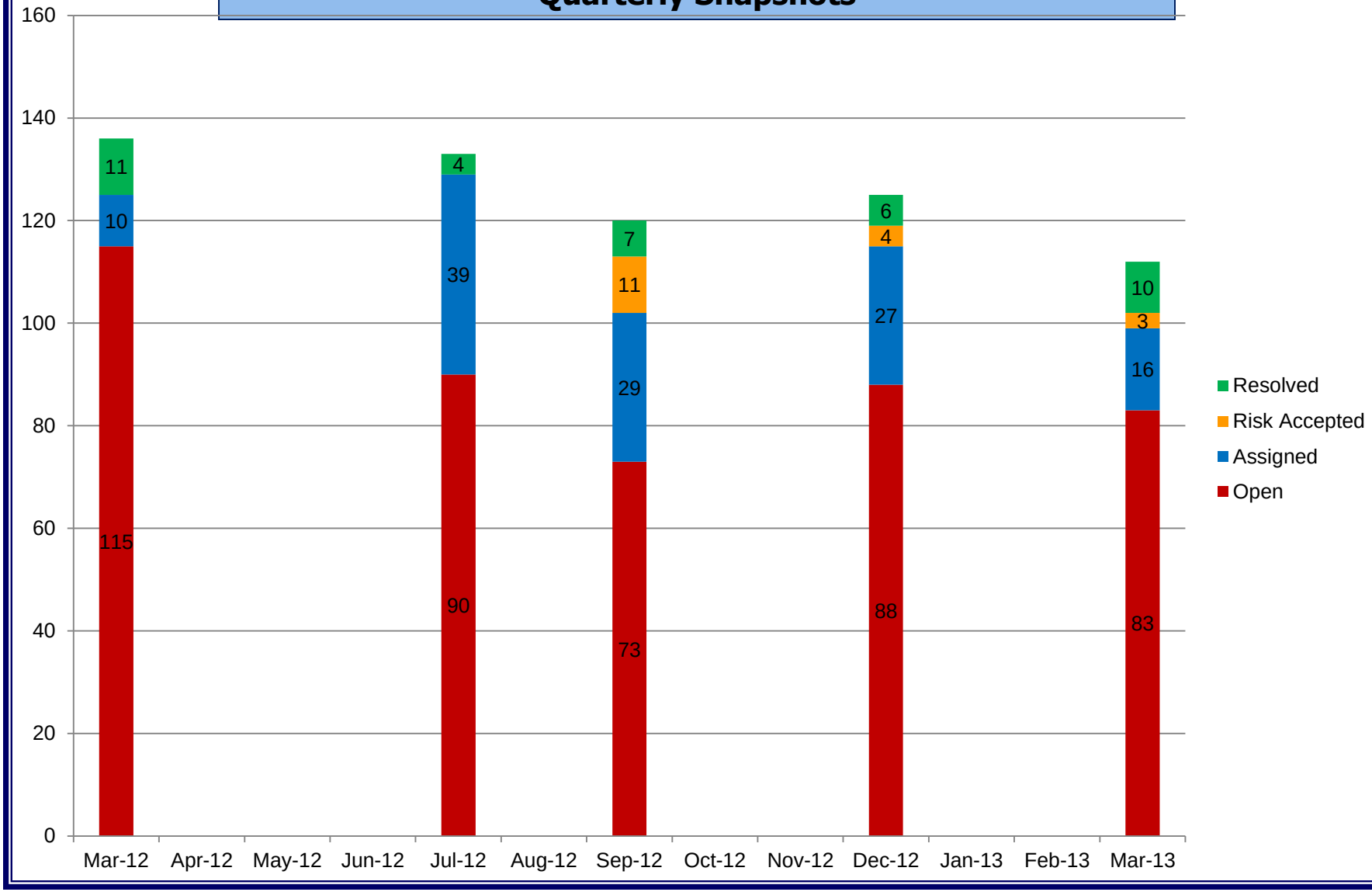
Next we review Metrics

Questions regarding High-Profile Problems?

Q1 2013	
New	8
Resolved	10
Risk Accepted	3

All of 2012	
New	46
Resolved	32
Risk Accepted	14

Quarterly Snapshots

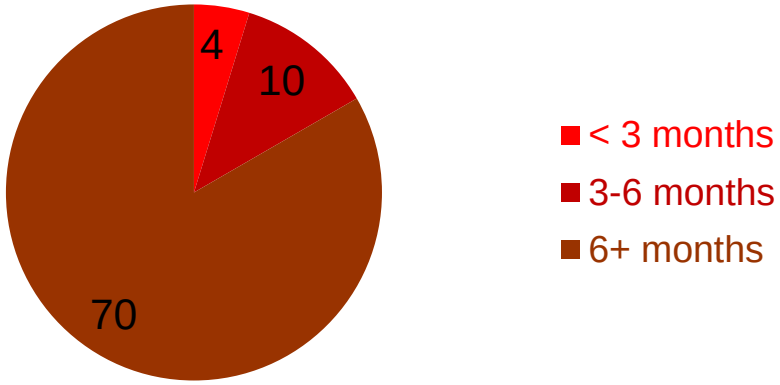


Resolved — Fixed + Closed
Risk Accepted — We intend to live with this
Assigned — Tech is working on it
Open — Inert

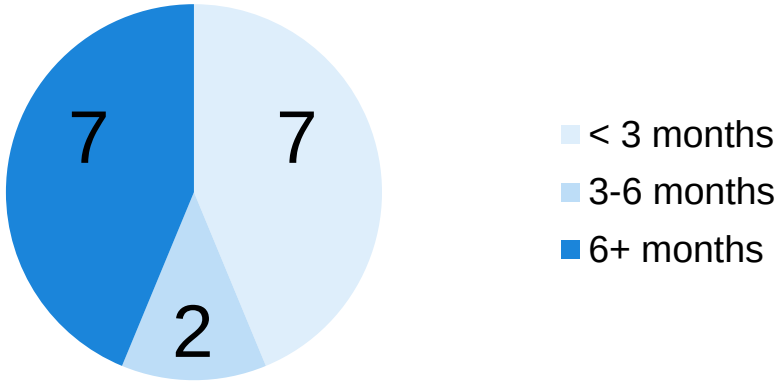
We are making progress in shrinking the Queue
*And in more accurately distinguishing between **Assigned** and **Open***

Age of Problems Today

Open by Age



Assigned by Age



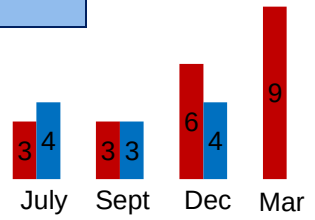
The Supervisors are confident that their staff are working on Assigned Problems
… but since Problems take back seat to other responsibilities, progress is slow

Mega-Problem Status

Tungsten Stumbles

Owner + Implementer = InfraOps + InfraOps

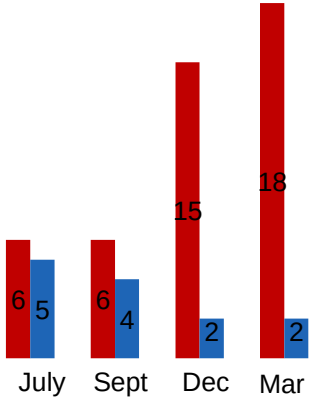
Roadmap = Data & Storage



Data Center Redundancy Degrading

Owner + Implementer = InfraOps + InfraOps

Roadmaps = Data & Storage, Server, Core Transport



Servers Running Unsupported OS and Applications

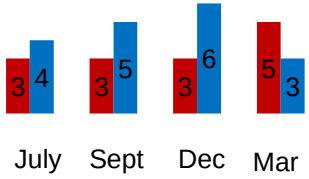
Owner + Implementer = Security + Numerous

Roadmaps = Application, Security, Server

Security Patching

Owner + Implementer = Security + InfraOps

Roadmaps = Application, Security, Server



Blue = Assigned

Red = Open

- Blue has been shrinking: We are becoming more accurate about Assigning only when a tech actually has cycles available
- ***Data Center Redundancy Degrading*** has been degrading, as we discover additional interlocking Problems

Next Steps

What action items have we generated?

Who owns them?

Recall that Mega-Problems are composed of Problems with domino relationships … such that if one Problem is triggered, that may in turn trigger its siblings, a synergy which creates an Incident quite a bit larger than the scope of the component Problems would imply … a 'Normal Accident", to use the lingo

Future

What would you like to see in the June Deep Dive?

1. Status of Existing P1-P2
2. Review of New P1-P2
3. Focus topic: RCAs -- How to Make Them Go Faster
4. Metrics

More of, less of, keep the same, something different?

Andy

Dirk

Joe

Mary

Ron

Sonja

Tony

To meet your particular needs in June, what would you like to see done:
The same? Differently?

Appendix

Stuart's Summary

(A) ***Tungsten Stumbles***

Tungsten experiences on-going issues, sometimes service-affecting

(B) ***Data Center Redundancy Degrading***

Portions of the *deep infrastructure* are degrading

- Highly-Available becoming Highly-Unavailable
- Data Centers vulnerable to complex disruptions
- Interferes with capacity upgrades, bug fixes, security patching

(C) ***Servers Running Unsupported Applications and OS***

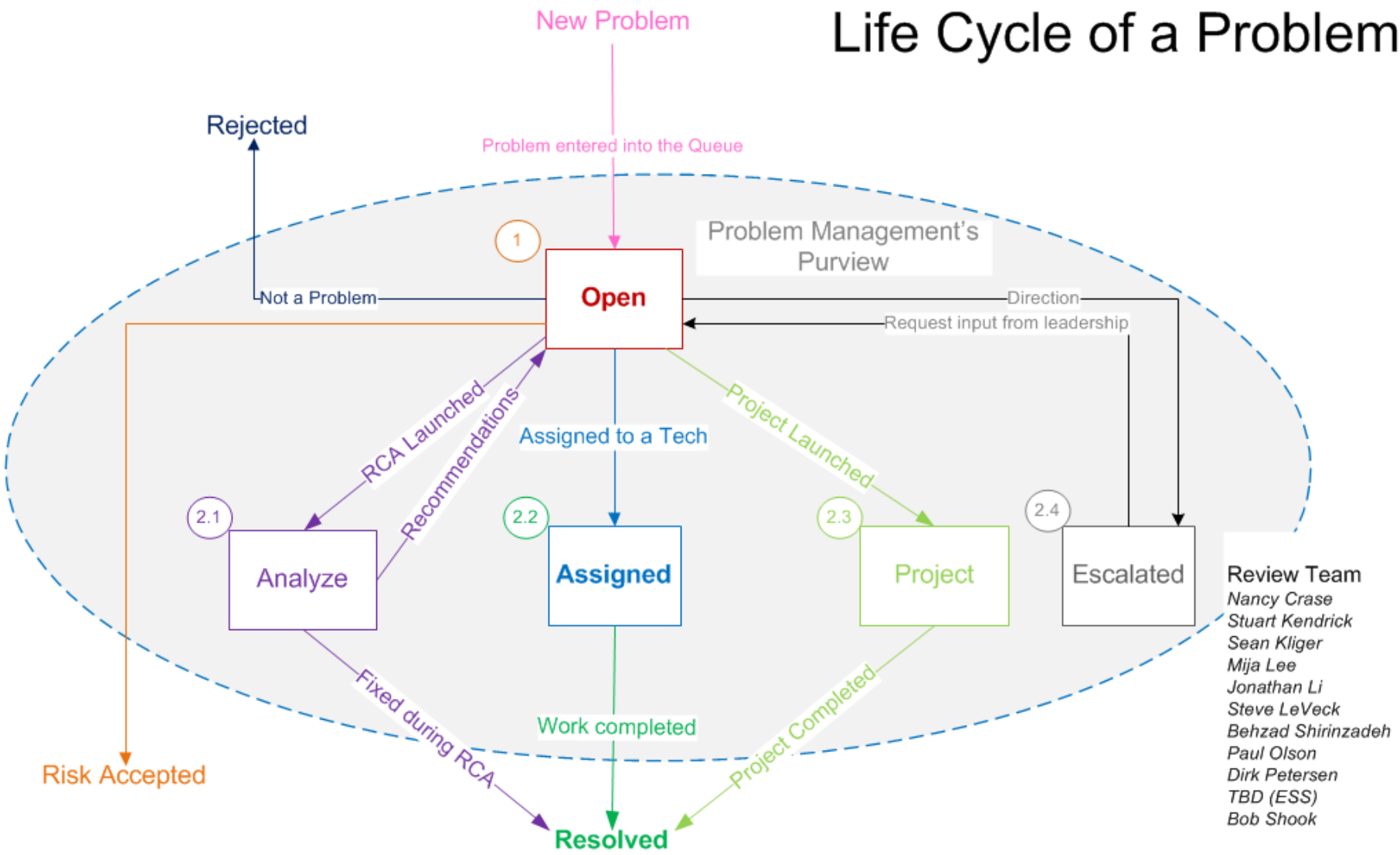
Aging applications living on aging servers

- Hard to identify owner
- Exposed to security vulnerabilities
- General ignorance and age suggests *unknown unknowns*

(D) ***Security Patching***

Steadily falling behind

Life Cycle of a Problem



Legend

Problem Management States

Open	Inert: Stalled on resources
Analyze	Root Cause Analysis team instantiated
Assigned	Tech is working on the Problem
Project	Project instantiated
Escalated	Disagree, seeking leadership direction

Problem Management Sources/Destinations

New Problem	From Incident Mgmt, Techs, Problem Manager
Rejected	Not a Problem
Risk Accepted	We intend to live with this
Resolved	Fixed + Closed

skendric 2013-03-19

Priority Matrix

Priority is the relative urgency of the problem , calculated using the look-up table below

Rating	Description
1	Most important
2	
3	Average importance
4	
5	Less important

		Impact				
		1	**2**	**3**	**4**	**5**
Likelihood	**1**	1	1	2	3	4
	2	1	2	2	3	4
	3	2	2	3	4	4
	4	3	3	4	5	5
	5	4	4	4	5	5

Likelihood of 1 = Likely to occur at least once over the next year
Likelihood of 5 = Unlikely to occur at all over the next year