Problem Management Lite
Quarterly Deep Dive

Agenda

➢**Objectives:**

❑ RCA Review	5 minutes	**1:05 – 1:10**
❑ Mega-Problem Review	5 minutes	**1:10 – 1:15**
❑ Priority 1 & 2 Review	50 minutes	**1:15 – 2:05**
❑ Next Steps	20 minutes	**2:10 – 2:25**

➢**Goals:**

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

December 19, 2012

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

RCA: A Year in Review

Five RCAs

1.	Q4 2011	Rhino RCA Phase I
2.	Q1 2012	Tungsten RCA Phase I
3.	Q2 2012	Tungsten RCA Phase II
4.	Q2-Q3 2012	Rhino RCA Phase II

Results

1.	Found and fixed two issues	*Success*
2.	Identified handful of issues	*Success*
3.	Identified next steps	*Iced*
4.	Mitigated third issue	*Risk Accepted*

Inhibitors

Typically, each RCA stumbles during its first month:

1. Free up staff
2. Expand permissions to include RCA techs
3. Hack together monitoring & analysis tools

And thereafter picks up steam.

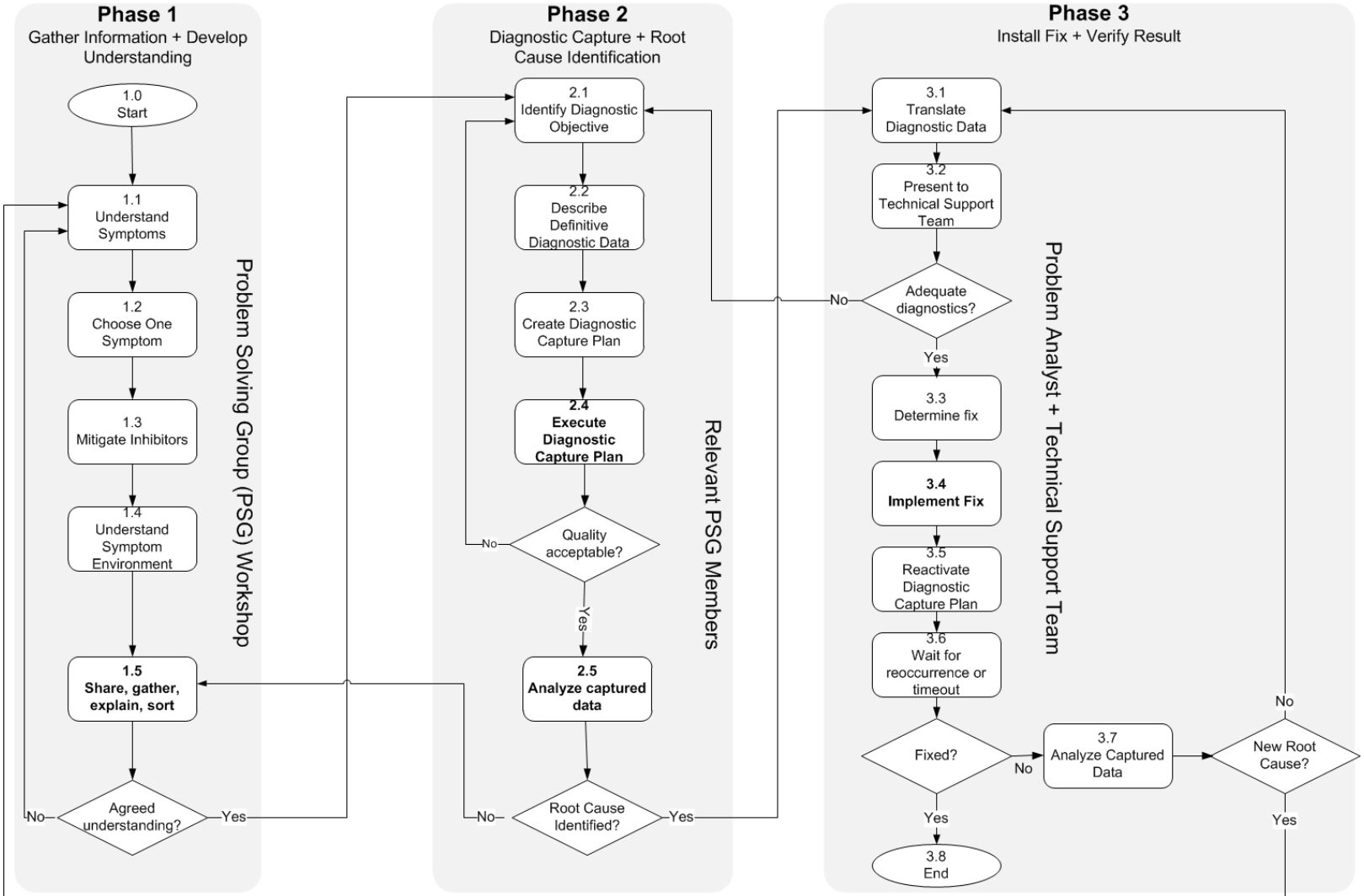
Future

Kick-off each RCA with a *Mitigate Inhibitors* meeting with leadership, to clarify resource, permissions, and tools constraints

Develop a standardized methodology which adapts to Inhibitors

Refresh toolkit

RCA Methodology



Transition

Next we review Mega-Problems

Questions regarding RCAs?

Criteria for a Mega-Problem

The Bucket

The Bucket is structured for techs and their Supervisors, chunked into mouthfulls which make sense at this level, typically:

- Contained within a single technology group and ideally
- Doable by a single tech

More work needed to implement this fully

Mega-Problems

For leadership, I maintain a separate *Mega-Problem* queue, abstracted from *The Bucket* -- these are the items which I believe are worth leadership attention.

Historically, *Mega-Problems*, and only *Mega-Problems*, are what you saw during a Deep Dive.

Mega-Problems typically span technology groups and/or require Project-level intervention to address.

Criteria

To become a *Mega-Problem*, the issue must offer a noticeable … my judgment … chance of inflicting at least one of the following:

1. Disrupt services Center-wide

2. Disrupt our core business: *Grant/contract funded research*

3. Impact Center leadership (reputational issues)

Synergy

Most of these *Mega-Problems* encapsulate multiple *Bucket Entries*: typically when those entries interact synergistically: domino effect

One day, we may see a one-to-one mapping between a *Mega-Problem* and a *Bucket Entry* (aka a *Problem*)

… but this hasn't happened yet.

One-third of the issues in the Queue belong to a Mega-Problem

Mega-Problem Status

Tungsten Stumbles

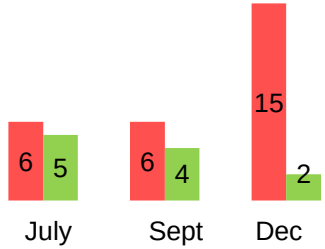
Owner + Implementer = InfraOps + InfraOps

Roadmap = Data & Storage

Data Center Redundancy Degrading

Owner + Implementer = InfraOps + InfraOps

Roadmaps = Data & Storage, Server, Core Transport



Servers Running Unsupported OS and Applications

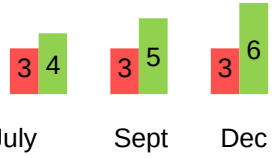
Owner + Implementer = Security + Numerous

Roadmaps = Application, Security, Server

Security Patching

Owner + Implementer = Security + InfraOps

Roadmaps = Application, Security, Server



Green = Assigned

Red = Open

Transition

Next we review the Priority 1 and 2 Problems

Questions regarding Mega-Problems?

{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 learn?
- ❑ Next steps

Tungsten-Cobalt Fumble

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

Incident Likelihood	Unknown — Three Incidents across three years: March 2011, January 2012, February 2012
Incident Impact	Widespread — Possible application disruption for hours; perhaps a few applications unavailable for days
Incident Recovery	High — ~200 CIT + NAG hours across several weeks, gradual restoration of end-user services*
Root Cause	Unknown

Summary

- Weekly: Tungsten logs errors about delays in reaching Cobalt
- Monthly: Fragile servers crash and require a reboot
- Three Incidents: Wide-spread disruption

- NetApp and HP tech support warn that these error messages indicate a serious problem

Status - Static

Analysis Plan	Analysis Effort - High
Spin up an RCA	30 – 200 hours 2-3 Staff

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

*Based on past experience

Tungsten Performance Degraded

Priority	Status	Owner	Service Impacted	Customer Impacted	Start
2	Open	Architecture	Data Storage	FHCRC & SCCA	Oct 2011

Incident Likelihood	Ongoing
Incident Impact	Widespread — Slow response from applications
Incident Recovery	Low — Users restart applications if they time out
Root Cause	Known — Load

Summary

Load has increased to the point where application performance suffers noticeably

Status – Degrading

Historically, we have deployed various measures to mitigate Tungsten OS overload, including:

- Giving up high-availability (migrating from Active/Standby to Active/Active to employ both Heads)
- Buying additional NetApps (Thorium for SCHARP, Copper for SciComp)
- Application optimizations

Mitigation Plan	Mitigation Effort - Low
Spin up Project to identify options	15 hours 2 staff

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

SQL Server Crumps During Storage Events

Priority	Status	Owner	Service Impacted	Customer Impacted	Start
2	Open	InfraOps	Database Hosting Enterprise App Mgmt	FHCRC + SCCA	2008

Incident Likelihood	Unknown — Unplanned events interrupting Carbon, Tungsten, or the J4-401 Network would trigger this
Incident Impact	Widespread — Enterprise SQL and MIS Applications unavailable, possible data loss
Incident Recovery	High — ~50-100 CIT hours across a week, gradual restoration of end-user services*
Root Cause	Known — Software bugs + configuration errors

Summary

- Our various SQL Servers do not tolerate minor bumps in access to Storage, unmounting their databases abruptly, sometimes suffering database corruption
- As a work-around, we always shut down SQL Server whenever we anticipate Storage or Network maintenance
- Naturally, this work-around is only effective for planned events

Status - Static

Resolution Plan	Resolution Effort - Medium
1. Use the Development environment to test fixes 2. Use a maintenance window to verify that the fixes work	1. 30 hours 2. 30 hours 1 staff

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

*Based on past experience

Data Centers Intermittently Isolated During Router Failure

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

Incident Likelihood	Low
Incident Impact	Widespread — Brief isolation of a Building plus its Data Centers; Possible data loss and application corruption
Incident Recovery	High — ~20-200 CIT + NAG hours across one week, gradual restoration of end-user services
Root Cause	Unknown

Summary

- By design, our Buildings can survive the loss of either of their Redundant Routers
- However, during routine validation, we discovered that portions of a Building became isolated during the loss of a Router
- The highest profile example has been the DF-120 Data Center (Data Protection)

Status - Static

Analysis Plan	Analysis Effort - Low
During a maintenance window, replicate the problem, analyze	20 hours 2 staff

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

Uncertain Router Failover for NetApps

Priority	Status	Owner	Service Impacted	Customer Impacted	Start
2	Open	InfraOps	Data Storage	FHCRC	Dec 2009

Incident Likelihood	Low — Loss of a Redundant Router or Switch might trigger this
Incident Impact	Widespread — Isolation from Storage for Clients located outside a given Building
Incident Recovery	Medium — ~2-50 CIT + NAG hours across one week, gradual restoration of end-user services
Root Cause	Known — Configuration error

Summary

- By default, our NetApps are not configured to take advantage of our Redundant Router scheme
- This is surprising, as we use one of several popular and standard Redundant Router schemes
- In contrast, out-of-the-box Windows, Linux, and Mac OS hosts are configured suitably
- Nevertheless, our analysis says that NetApps ship by default with a configuration choice ('FastPath') which dodges our scheme
- This issue only affects Clients located outside a given Building (e.g. home drive users of Tungsten located outside Yale)

Status - Static

Analysis Plan	Analysis Effort – Low
• Make the configuration change on a guinea pig NetApp • Reboot the associated Redundant Router • Analyze results	10 hours 2-3 staff

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

vColo Does Not Reliably Recover from Storage Bumps

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 - 20 CIT + NAG hours across one day, gradual restoration of end-user services
Root Cause	Unknown

Summary

- Sometimes when Storage stumbles, some vColo Guests (and vColo mgmt functions) require reboots to recover
- During pathological Storage Incidents, this is unsurprising
- But it happens even during clean + fast Storage head fail-overs, when other Storage clients recover without issue

Status - Static

Analysis Plan	Analysis Effort – Medium
Spin up RCA	200 hours 3 staff

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

Uncertain Router High-Availability

Priority	Status	Owner	Service Impacted	Customer Impacted	Start
2	Open	InfraOps	Connectivity	FHCRC + SCCA	April 2011

Incident Likelihood	Unknown — Loss of a Redundant Router or Switch might trigger this
Incident Impact	High — Isolation of a Building plus its Data Centers from the rest of Campus; Possible data loss and application corruption
Incident Recovery	High — ~20-200 CIT + NAG hours across one week, gradual restoration of end-user services
Root Cause	Known — Software bugs + configuration errors

Summary

We are not confident that the Data Center, Building, and Campus Core switches/routers are still highly-available (HA)

Status - Degrading

We are stalled on patching Data Center switches in part because we fear triggering this issue

Resolution Plan	Resolution Effort – Low to Medium
1. Characterize current state, Resolve current issues	1. 12 hours / 1 staff
2. Schedule quarterly Building + Campus validation	2. 8 hours / quarter
3. Pick one Data Center per quarter, validate behavior	3. 100 hours / quarter

- ❑ Do you understand the risk? If not, what do you need to learn?
- ❑ Next steps: Assign | 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), requiring manual intervention and widening the gaps

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 learn?
- ❑ 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
- During the recent J4-401 Ethernet switch chassis upgrade, *Audhumla* was isolated (it has only one NIC) and Security noticed (calls to x5000 were receiving 'Fast Busy')

Status - Static

- This is a design decision we made years ago – we weren't willing to spend the $$ to buy a second one
- The proposal to add a second *Audhumla* did not make the FY'13 funding cut
- Loss of an Ethernet switch, *Audhumla's* single NIC, or all of *Audhumla* itself would result in 'Fast Busy' to x5000 and x5700
- I'm using Problem Management to verify that this audience knows and accepts this risk

Resolution Plan	Resolution Effort - Low
Purchase a second Audhumla Configure it as a hot-standby	$xx capital 1-2 staff / yy hours

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

Intermittent Failed Network Connections

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 learn?
- ❑ Next steps: Leave Open | RCA | Risk Accept …

Transition

Next we review Metrics

Questions regarding P1s & P2s?

Quarterly Snapshots

March 30		July 5		Sept 25		Dec 20	
New	34	New	9	New	5	New	9
Open	81	Open	72	Open	68	Open	79
Assigned	9	Assigned	39	Assigned	29	Assigned	27
Project	0	Project	0	Project	0	Project	0
RCA	1	RCA	0	RCA	0	RCA	0
Escalated	0	Escalated	9	Escalated	0	Escalated	0
Total Active	91	Total Active	120	Total Active	97	Total	106
Risk Accepted	0	Risk Accepted	0	Risk Accepted	11	Risk Accepted	4
Rejected	0	Rejected	0	Rejected	5	Rejected	0
Resolved	11	Resolved	4	Resolved	7	Resolved	6
Total Closed	11	Total Closed	4	Total Closed	23	Total Closed	10

Problem Status

Assigned	Tech is working on the Problem
Escalated	Disagree on Status, seeking leadership direction
Open	Inert: not working on it
Project	Project instantiated
RCA	Root Cause Analysis team instantiated
Risk Accepted	We intend to live with this
Rejected	Not a Problem
Resolved	Fixed + Closed

Next Steps

What action items have we generated?

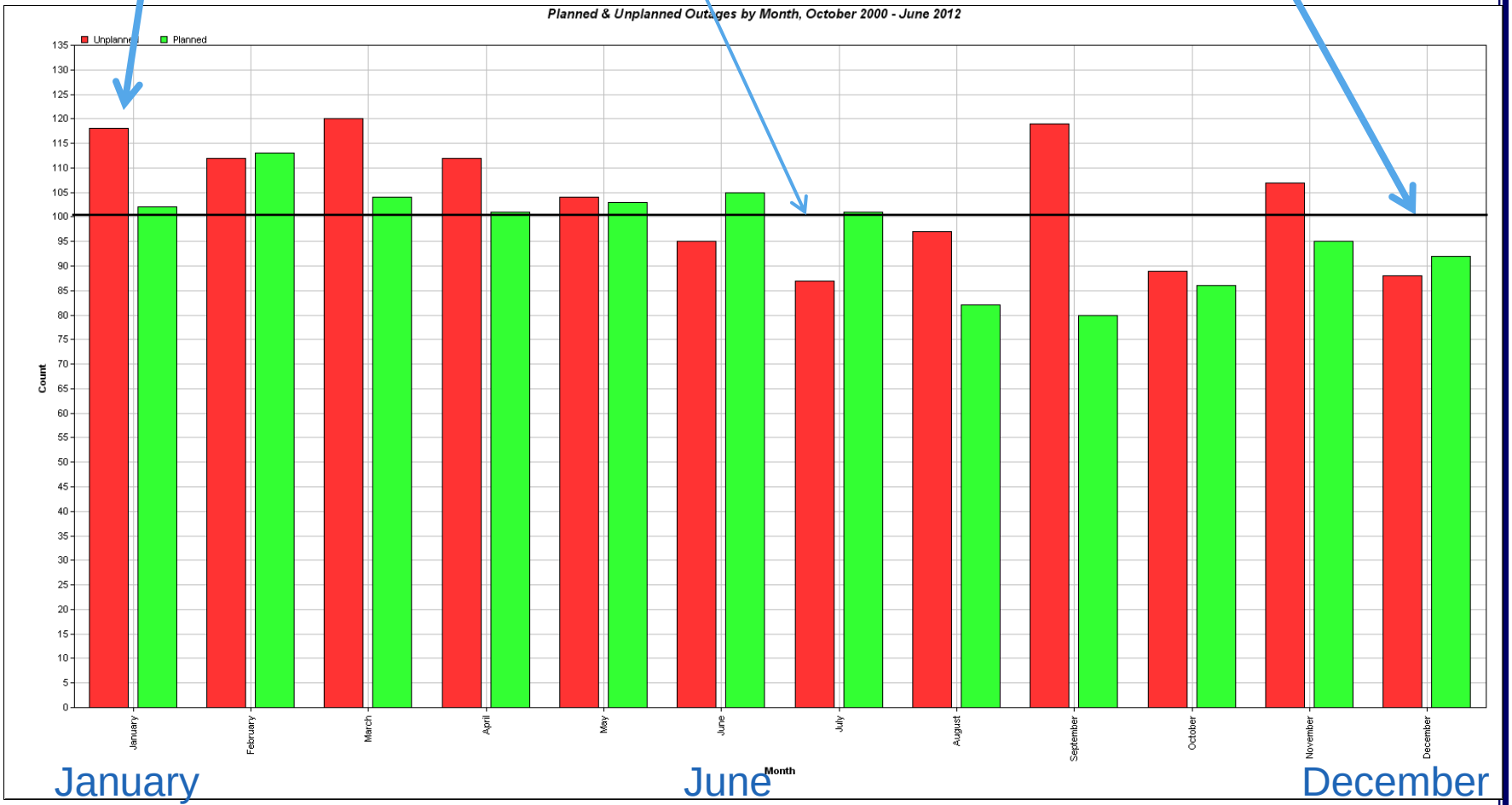
Who owns them?

Planned & Unplanned Outages by Month, October 2000 – June 2012

Coming Soon

Line marking '100'

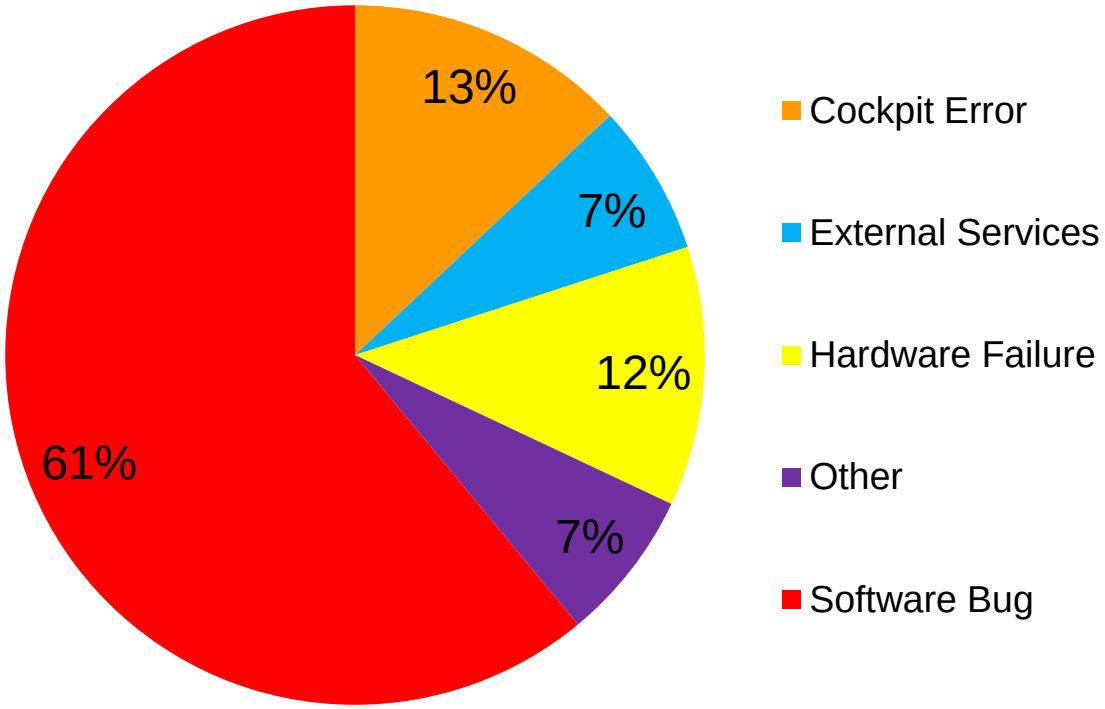
You are Here



January June December

Past performance does not guarantee future returns …

Causes of Unplanned Outages: October 2000 – June 2012



Would regular patching shrink Software Bug?

http://www.skendric.com/problem/incident-analysis/2012-06-30/What-Takes-Us-Down.pdf

Future

Is this what you want from a Prob Mgmt Deep Dive?

Goals:

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

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

1. Review P1s & P2s
2. Metrics
3. Process Engineering

Appendix

Stuart's Summary

(A) Consolidated Storage experiences on-going issues

(B) 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) Aging applications living on aging servers

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

(D) Limited visibility into our flock of Problems: *Unknown unknowns*

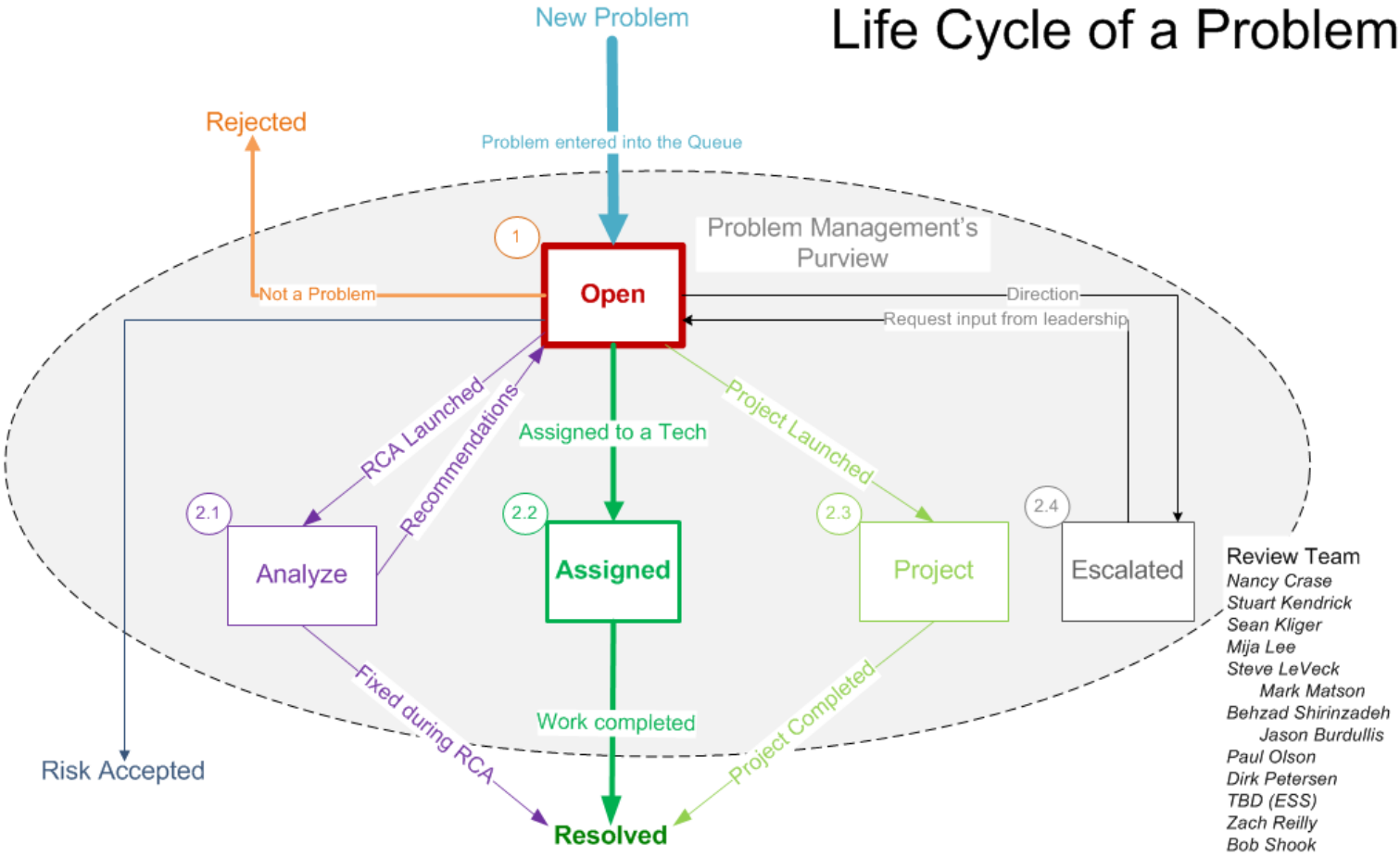
Predicting is hard, particularly the future –Niels Bohr

Process Engineering

1. Do we want to track time spent fixing Problems? If so, how?
2. How do we prioritize fixing Problems into our Maintain / Build / Improve work?
3. How do we track progress toward resolution?
4. What value are we delivering with Problem Mgmt?

Action item: Joan to pursue ‘pooled’ resource model

Life Cycle of a Problem



Legend

Problem Management States	
Open	Inert: Not yet reviewed or 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, Project Managers ...
Rejected	Not a Problem
Risk Accepted	We intend to live with this
Resolved	Fixed + Closed

skendric 2012-08-16

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