Problem Management Lite
Quarterly Deep Dive

Agenda

➢**Objectives:**

❑ Review Process	10 minutes	**10:35 – 10:45**
❑ Metrics	10 minutes	**10:45 – 10:55**
❑ Visit Meta-Problems	40 minutes	**10:55 – 11:35**
❑ Review & Summary	10 minutes	**11:35 – 11:45**
❑ Next Steps	5 minutes	**11:45 – 11:50**

➢**Goals:**

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

July 10, 2012

Stuart Kendrick

Familiarize yourself with the deck's resources by scanning the Appendix (p.22)

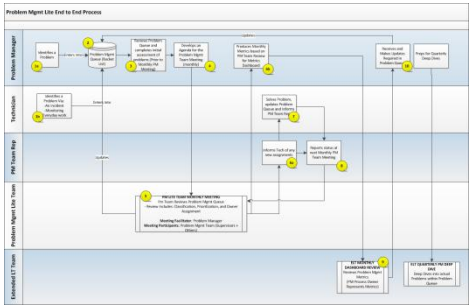
FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Problem Management Evolution

Past
Reactive Approach

- Multiple Tools in Use
- Limited leadership visibility
- No unified view
- No review process
- No entry criteria, no prioritization
- Duplicates and repeats

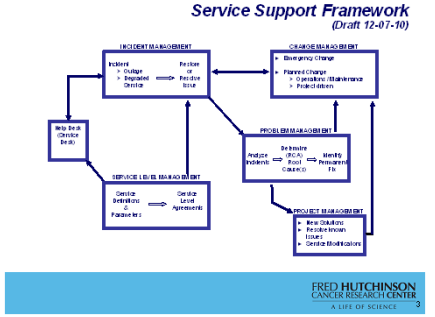
Today
Problem Mgmt Lite



Problem Mgmt Lite: leverage existing staff & procedures to deliver value while we address higher priority Processes

- One list
- Visibility to leadership
- Consistent reporting
- Monthly + quarterly review
- Defined Categories & Priorities
- Accountability

Future
Integrated Prob Mgmt



- Deepen maturity
- Assess relationship with Service Support Framework
- Review priority wrt other initiatives
- Formalize linkage with Incident Mgmt and Dashboard Review
- Extend beyond CIT

abozzuti

Process Definitions

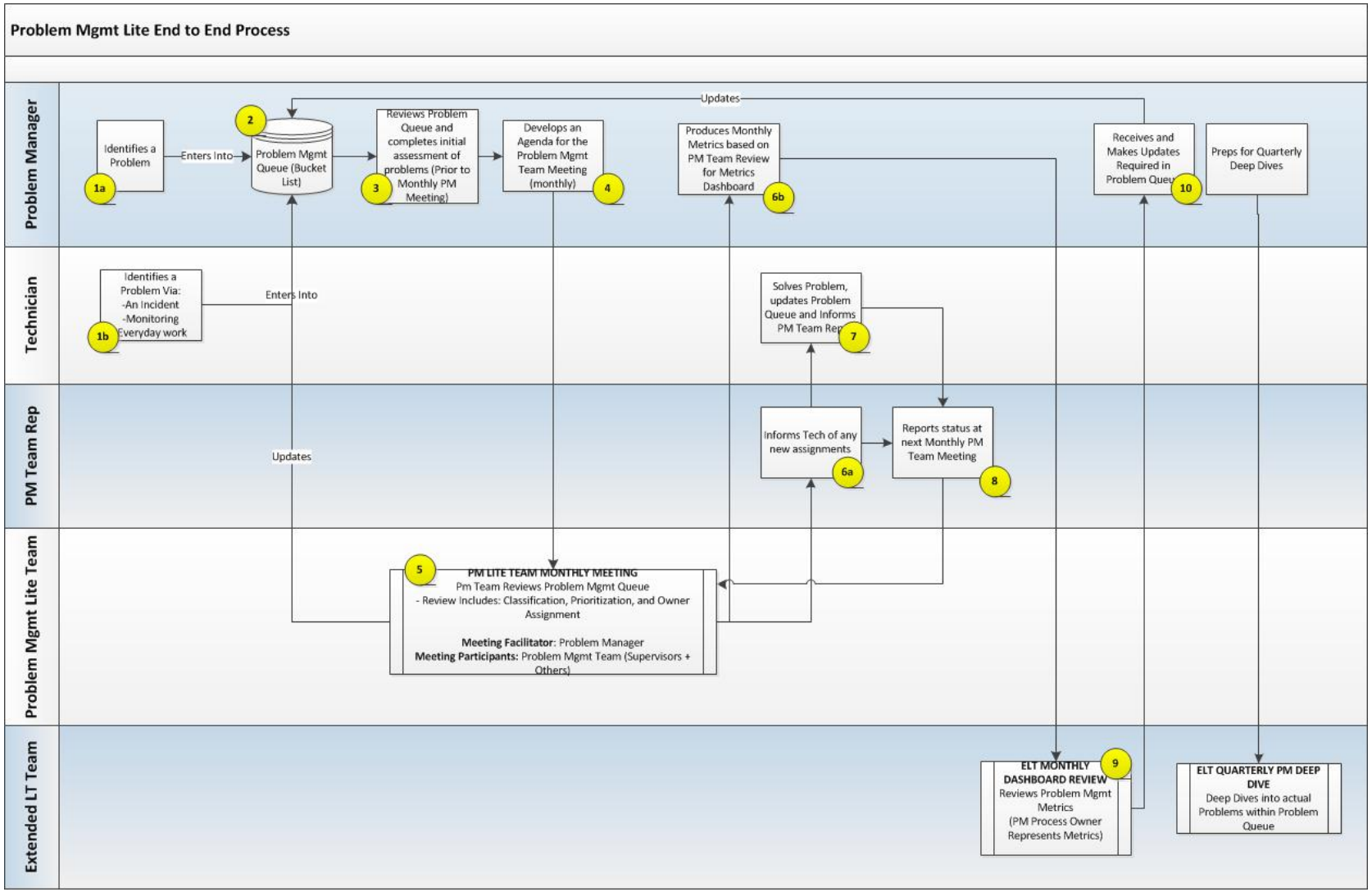
Problem A cause, or potential cause, of an incident that has already, or may in the future, interfere with a defined IT service

Priority Derived from a matrix of subjective *Impact* and *Likelihood* ratings

Impact Captures the cost to the business should this issue bite us (1 = High Impact … 5 = Low Impact)

Likelihood Captures the chances of the issue biting us over the next year (1 = Likely … 5 = Unlikely)

Review the Process

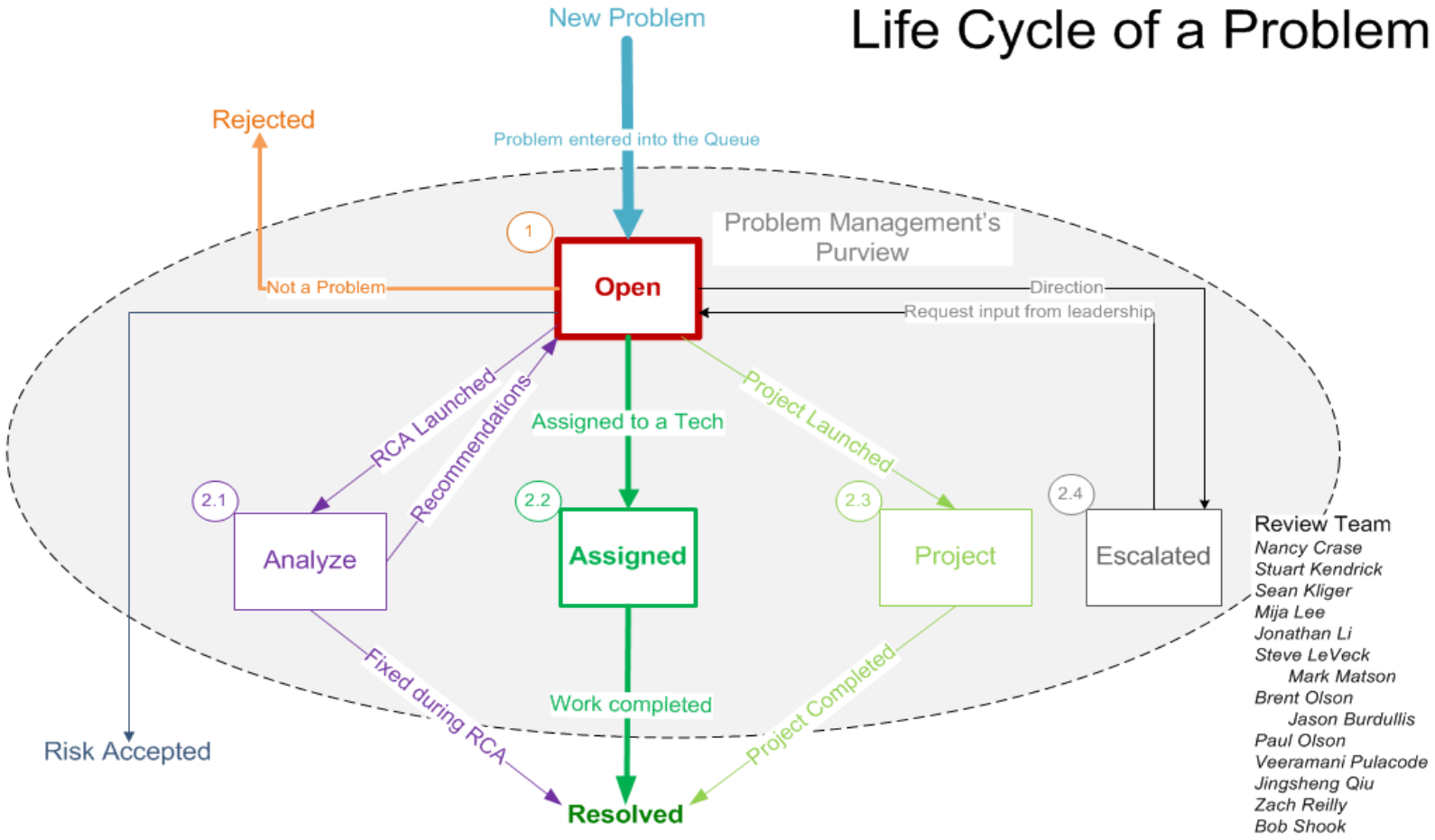


- Techs enter issues, Supervisors review monthly, ELT reviews quarterly
- Additional visibility and feedback via Dashboard Review

ssgrover

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

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 2012-05-30

Currently, we dump into Escalated any Problems which we want to Reject or Risk Accept, in addition to Problems about which we disagree

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Process Successes

Successes in 2012

- ✓ Review Team has met monthly
- ✓ Meeting agenda, notes, and discussion posted to http://lists.fhcrc.org
- ✓ Team discusses ways to improve process during Review Meetings
- ✓ Owner and Lead refining Process & Metrics Reporting
- ✓ We have categorized 61 New Problems, Closed 20 (Resolved)

Closed in 2012

- ✓ Carbon Client Resetting LUNs
- ✓ TCP Window Limiting Throughput on Fred
- ✓ Charon Denied Some Telecommuter VPN Connections
- ✓ High Mortality Rate for Robert Bradley HPC Jobs
- ✓ Parking Server Running on Aging Hardware
- ✓ Dining Server Running on Aging Hardware
- ✓ Charon Vulnerable to DoS Attack
- ✓ Temperature Sensors on NetApps Failed
- ✓ M4 NetApp still in production
- ✓ Dell management agent not installed on new vColo hosts

Process Challenges

1. Visibility

In 2012, I have been the source of 90+% of the incoming entries; my visibility is limited to portions of InfraOps and SciComp

Risk ➔ Limited visibility into Problems

Mitigation ➔ Stuart continues to evangelize Process with peers

2. Team Cohesion

Team struggles to apply definition of Problem to new entries

Risk ➔ Consumes monthly meeting time

Mitigation ➔ Review Escalated in future Deep Dive, ask leadership for direction

3. Process Immaturity

Team members spend 3-4 hours/month in review meetings (instead of original 1 hour/month)

I spend 25% of my time managing the process

I want to stall on addressing these issues: get through our first Deep Dive, a future review of the Escalated items, and the onboarding of the new SOPS Supervisor – see what happens when we have a few more months under our belt.

I will try to limit discussion here, wanting to reach the Problems

Questions about Process?

Next will be Metrics

Questions about Process?

Quarterly Snapshot

Last Quarter
March 30, 2012

Open	81
Assigned	9
Project	0
RCA	1
Escalated	0
Total	91

Delta

New	21
Resolved	3
Rejected	1
Risk Accepted	3
Delta	28

This Quarter
June 30, 2012

Open	72
Assigned	39
Project	0
RCA	0
Escalated	9
Total	120

Problem Status

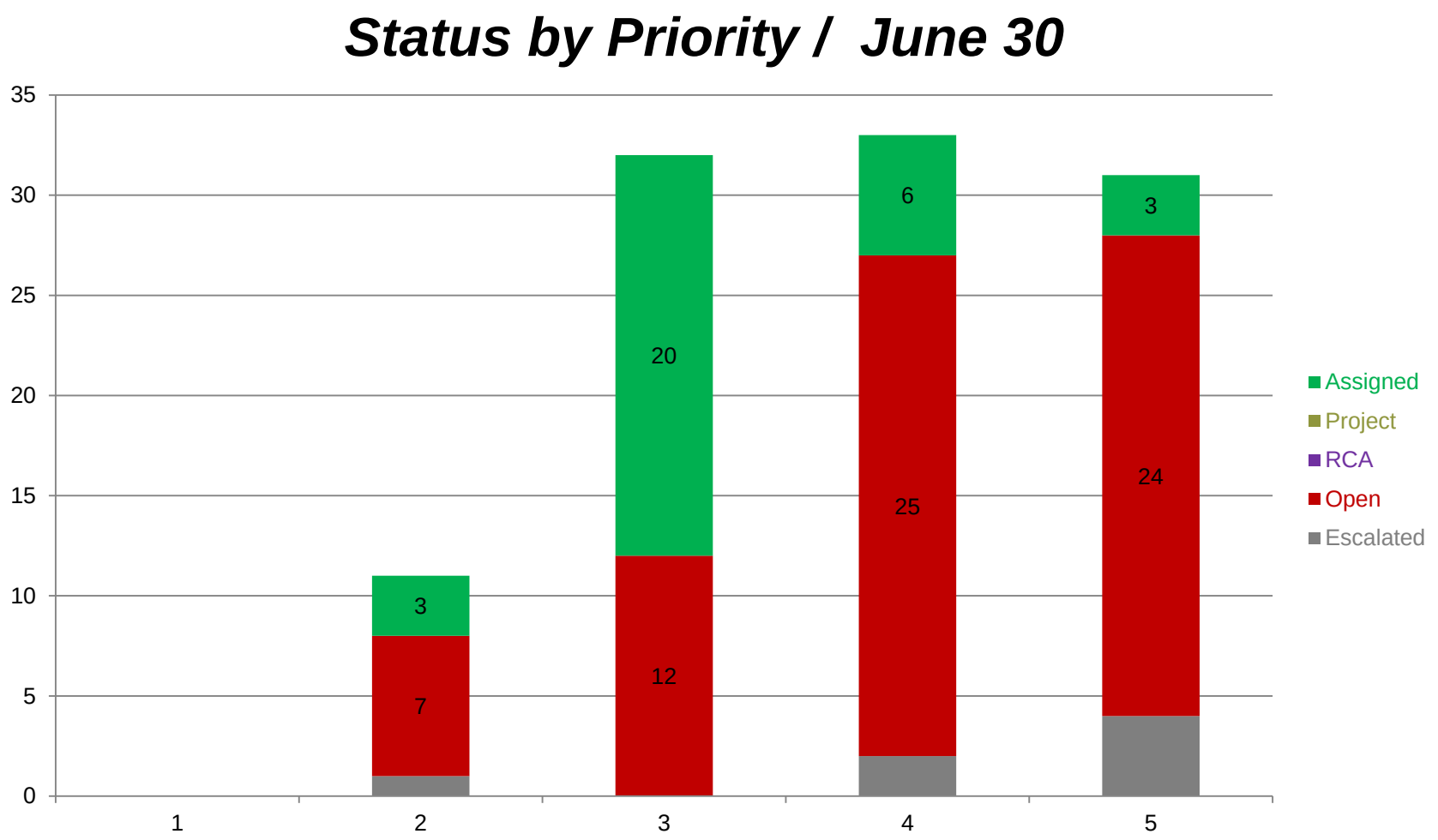
Assigned	Tech is working on the Problem
Escalated	Disagree on Status, seeking leadership direction
Open	Inert: not working on it
Project	Project instantiated
RCA	Root Cause Analysis team instantiated
Risk Accepted	We intend to live with this
Rejected	Not a Problem
Resolved	Fixed + Closed

Coming to a Deep Dive in your Future:
Age of Problems

Changes from last quarter
- More Problems
- Migration from Open to Assigned

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Metrics – How Many Problems by Priority?

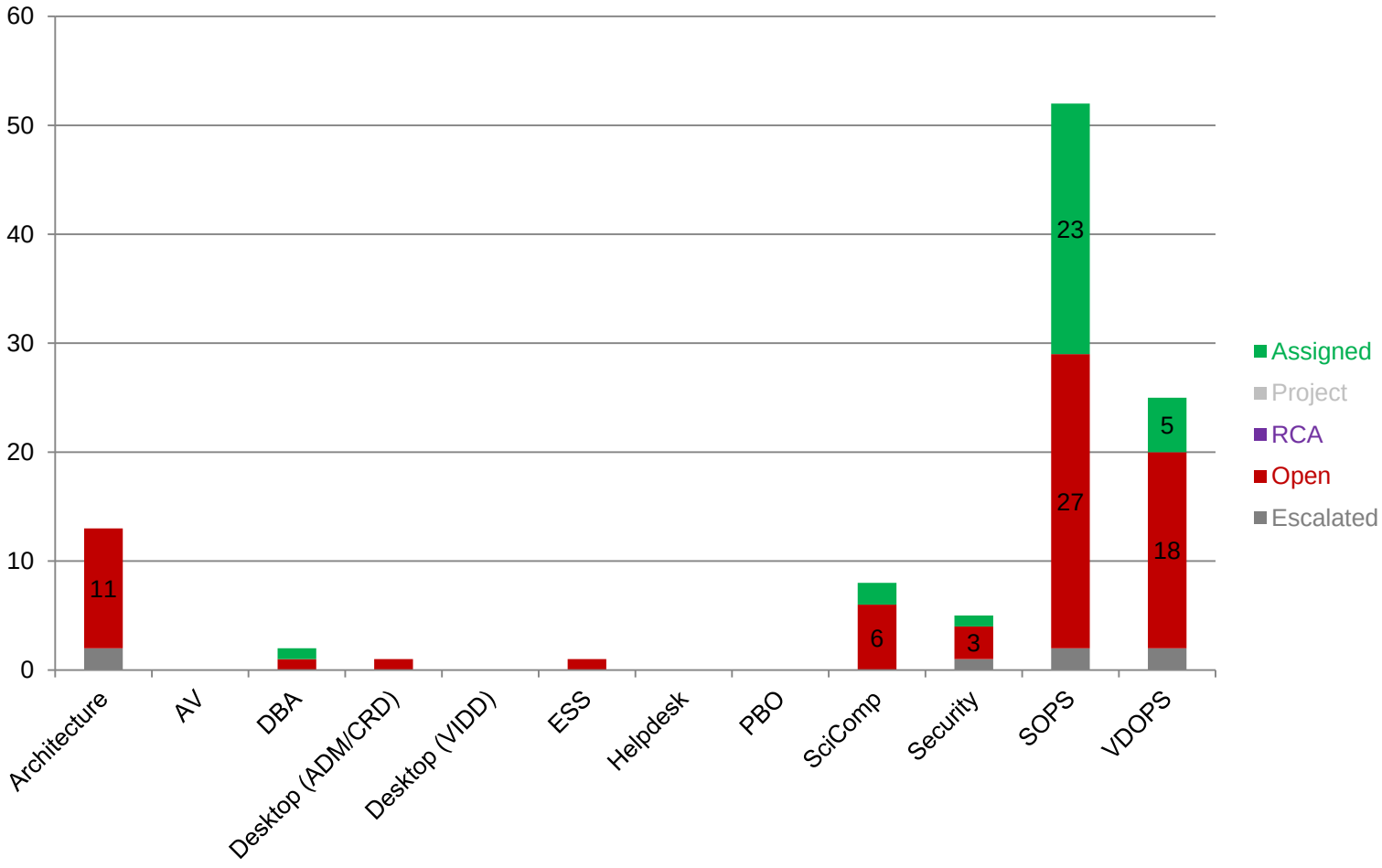


- Two-thirds of our Problems are Open
- We're being conservative about P1 – none yet

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Metrics – Who Owns our Problems?

Status by Department / June 30

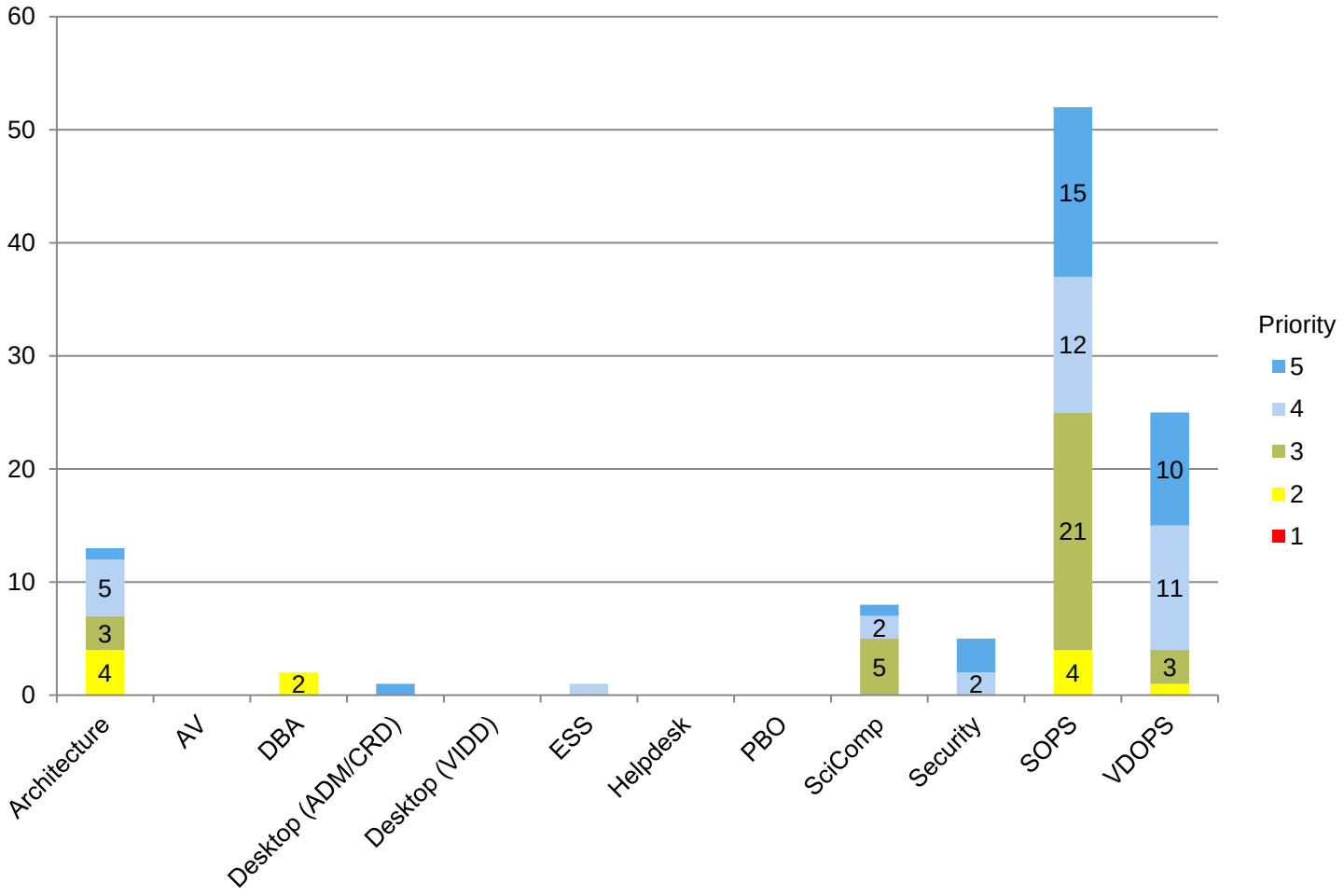


SOPS owns most of our Problems, with VDOPS trailing

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Metrics – Who Owns What Priorities?

Priority by Dept / June 30



Highest priority Problems live in *InfraOps* and *Architecture*

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Questions about Metrics?

We are about to visit Problems

Questions about Metrics?

High Impact Meta-Problems

The Problems as listed in the Queue are named and organized to meet the needs of hands-on staff.

Not suitable for this audience.

For this forum, I have bundled related Problems, to meet three goals:

- Cater to a higher level audience
- Conserve our time
- Emphasize the toxic synergy between Problem entries

I will skim through these five (5) Meta-Problems, then offer you a chance to dig deeper

Red = High Impact

Purple = Medium Impact

Orange = Low Impact

Grey = Unknown

These five (5) encapsulate most of the P1-P3s in the Queue

Goal

How do we engineer this process to make progress on fixing Problems?

1. Do we agree that this Meta-Problem is worth our time?
2. Who owns it?
3. How do we integrate solving this into our workload?
4. How do we track progress toward resolution?
5. Next steps

Please pull this slide out – we'll be referring to it

High Impact Meta-Problems

(1) Consolidated Storage Stumbles

History January 2010. *Intermittent minor Incidents plus 3 major Incidents, ongoing warnings*

Impact High: Disrupts Home & Shared Folders, vColo, Enterprise SQL Server, PeopleSoft …
Requires multi-day recovery

Root Cause Unknown

Recurrence Unknown

(2) Data Center Redundancy Degrading

History June 2009. Interferes with capacity upgrades, bug & security patching. *Multiple Incidents*

Impact High: Disrupts some, many, or all services within a Data Center
J4: Consolidated Storage, vColo, Enterprise SQL, PeopleSoft, Exchange 2003, Zimbra …
M1: COMPASS, Basic Sciences, HPC, SAGE, LabKey …
M2: HPC …
M3: SCHARP
M4: PHS, SciComp …
DF: Data Protection

Root Cause (a) Configuration errors, (b) Software bugs, (c) Unknown

Recurrence Unknown

Meta-Problems

(3) HPC Interactive Stalls

History August 2011. Partially mitigated during Rhino RCAs. *Daily events*

Impact Low: Researcher inefficiency and dissatisfaction

Root Cause Ten (10) known (but relative size of contributions uncertain)

Recurrence Ongoing

(4) Servers Running Unsupported OS and Applications

History July 2010. Numerous questions: we don't understand the exposure. *No Incidents*

Impact Medium + Unknown: Disruption to Clinical research functions plus Unknown

Root Cause Lack of Ownership

Recurrence Low

(5) Security Patching

History June 2010. *No Incidents*

Impact High: Intermittent disruption to Internet, to Transport (all IT services), to arbitrary services

Root Cause (a) Stalled patching process, (b) Stale firewall rule-set, (c) Highly-Unavailable services

Recurrence Low

HPC user frustration
Uncertain exposure around aging apps
Lagging on security patching

Pop Quiz

Goals for this Deep Dive

- Keep leadership apprised of current technical risks
- Identify next steps for highest priority Problems

1. How would you summarize our technical risks?

2. What action items has leadership assigned today and to whom?

I want to verify that I have conveyed the information I intended to communicate

Stuart's Summary

(A) Consolidated Storage and HPC experience 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*

Causes of Unplanned Outages (from Outages List 2001-2012)	
Cockpit Error	13%
Hardware Failure	12%
Miscellaneous	14%
Software Bug	61%

Predicting is hard, particularly the future –Niels Bohr

Fuzzy Risks

By not addressing these issues, we run the following risks:

1. They grow with load until a tipping point is reached, causing a major Incident
2. They create a fog which interferes with solving related problems
3. They constrain options, e.g. *I'd rather we didn't expand the use of system xyz because it is already creaking*
4. Users & techs are ground down by repeated, low-level static, creating long-term dissatisfaction

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

Objectives:

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

Meta-Problems: Keep ‘em or dump ‘em?

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

1. Review Metrics (including Age of Problems)
2. Visit highest impact Meta-Problems
3. Develop Prob Mgmt Process
4. Review Escalated Problems
5. Develop RCA Process
6. Review Outage Statistics

Appendix

Meeting Cadence

Who	What	When
ELT	Quarterly Problem Management Deep Dive	July 12
PM Review Team	Monthly Problem Queue Review	Aug 21
PM Review Team	Monthly Problem Queue Review	Sep 18
ELT	Quarterly Problem Management Deep Dive	Sep 27

Consolidated Storage Stumbles

Detail
- On three occasions, Tungsten lost touch with Cobalt for long enough to cause disruption
  March 22, 2011 *January 10, 2012* *February 1, 2012*
- Every few weeks/months, servers freeze because of Tungsten⇔Cobalt, requiring a reboot
- Tungsten logs severe warnings semi-weekly

Root Cause *Unknown*
Reoccurrence *Unknown*

RCA
- Offered the following options for finding Root Cause:
  #1 Install diagnostic capture tools ➔ Wait for next Stumble ➔ Analyze
  #2 Install diagnostic capture tools ➔ Induce a Stumble ➔ Analyze
- Leadership selected a limited version of #1
- We have not implemented

Risks (1) Repeat of previous Major Incidents, (2) Ongoing server freezes

Items		Priority
126	Cobalt Stumbles	2
384	Install NetApp Diagnostic Software	4
396	Move 3Par Diagnostic Software off Tungsten	4
401	Configure 3Par Diagnostic Software to retain 30 days of detailed data	4
403	Bugs in Cobalt OS	2
404	Bugs in Tungsten OS	2
433	Analyze Cobalt SCSI Logs	3

Repetition is the key to learning. --Jeffrey Case

Data Center Redundancy Degrading

Detail We buy two of things in an effort to provide High-Availability (HA)

Ethernet switches, Fibre Channel switches, Routers, NetApp heads, SQL Servers

- Chunks of Transport, Storage, and Clustered Servers are no longer redundant
- Some systems are Highly-Unavailable (HU): if either member fails, then the service breaks
- Causes include hardware failure, bugs, misconfigurations, and Unknown
- We have stopped periodic validation and bug & security patching

Root Cause *Software Bugs + Misconfiguration + Unknown* Reoccurrence *Unknown*

Risks
(1) A Data Center's services would become partly or entirely unavailable
(2) Recovery of failed systems can take days
(3) Large-scale outage required to augment capacity
(4) Large-scale outage required to perform routine patching (so we aren't)
(5) Interactions between HU components may cause surprising results

Items		Priority
24	Pokey Router Failover	2
101	Yale Routers Intermittently Unresponsive	2
122	iSCSI MPIO Misconfigurations	3
180	Bad Card in j4sr-b-esx	3
187	Uncertain Router Failover for NetApps	2
196	Misbehaving HA Ethernet NICs	3
199	Fred <=> Cobalt Path Errors	3
382	Tungsten HBA Ports Possibly Misconfigured	3
406	SQL Server Crumps During Storage Events	2
407	Pokey NetApp Failover	3
408	PeopleSoft Storage not HA	3

Normal Accident: small errors in judgment, flaws in technology, and trivial damage combine to form a large-scale event. –Charles Perrow

HPC Interactive Stalls

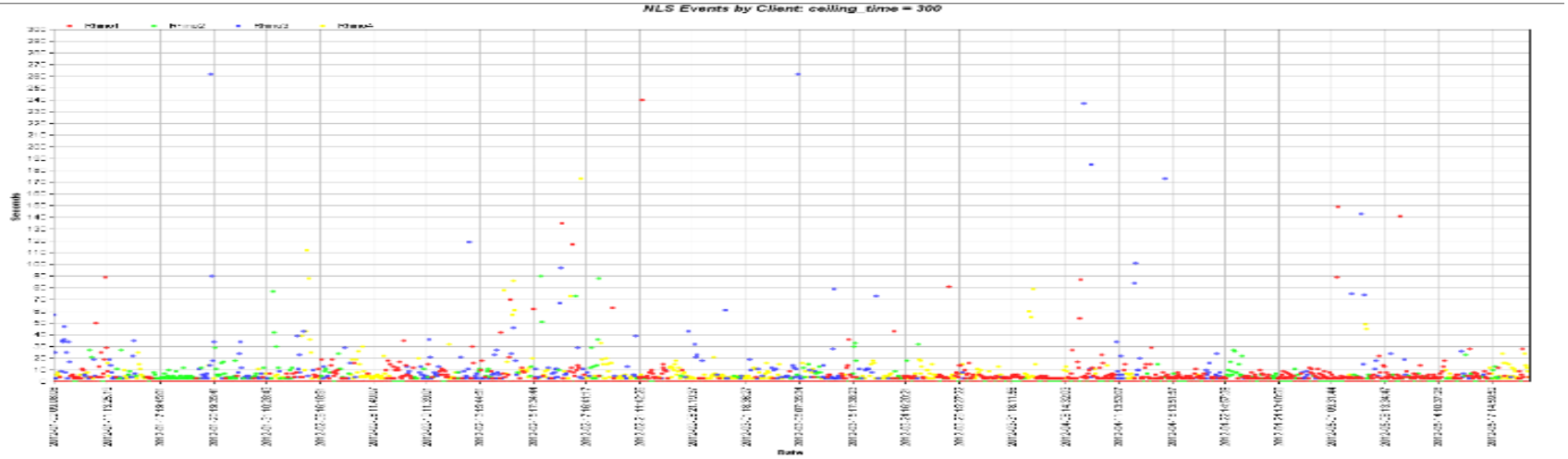
Detail HPC users experience intermittent interactive stalls lasting seconds, 10s of seconds and 100s of seconds.

RCA Identified 10 contributing factors. We picked one to mitigate, made a change in June, meet again on July 17 to review results and determine next steps

Root Cause *Multiple*
Reoccurrence *Unknown*

Risks (1) Researcher inefficiency and dissatisfaction

Items		Priority
	182 Reconfigure remaining concat volumes on Fred	4
	421 Interactive Stalls / Twink with dirty_xxx	3



FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Servers Run Unsupported Operating Systems

Detail A handful of servers run Windows NT (EoL December 2004) and Windows 2000 (EoL June 2010). Microsoft no longer ships patches: these servers are vulnerable to security issues. In addition, they run old versions of applications, similarly unsupported and unpatchable. Previous efforts to upgrade them have foundered. The application owners live within the Divisions. Questions to techs about these servers, applications, and users tend to elicit the response "I don't know": this suggests *unknown unknowns*.

Root Cause *Lack of Ownership, Age, Staff Resources*
Reoccurrence *Unknown*

Risks *xxx* stores [redacted]

xxx provides applications to CRD and Admin; compromise would knock out these apps:

FreezerWorks	Lab techs track which freezers hold which tissue specimens
Reference Manager	Researchers use this when writing papers for publication
Reflection	Clinical users employ this to reach research data
Thon	Development uses this during telephone fund-raising events
...	

...

Items		Priority
425	xxx Runs Unsupported OS	3
426	xxx Runs Unsupported OS	4
427	xxx Runs Unsupported OS	4
428	xxx Runs Unsupported OS	5
429	xxx Runs Unsupported OS	5

Security Patching

Detail
- We are months to years delayed on applying security patches to a range of systems, including Transport (switches & routers), Administrative Systems, and Research Systems
- An unknown number of these systems are open at the firewall, some deliberately, some in error

Root Cause *Staff Resources, Process*
Reoccurrence *Low*

Risks
(1) Loss of Partner and Internet connectivity during an external attack
(2) Partial or complete loss of Transport during an internal attack
(3) Loss or compromise to arbitrary services during internal or external attack

Items			Priority
	6	Validate Firewall Rule-Set (Close Accidental Holes)	4
	26	Revamp Radius Servers	4
	66	Standardize Windows Server Patching Procedure	3
	119	Vulnerability in TSM Clients	3
	144	Fix bugs in Catalyst 6500 OS	4
	409	DoS Vulnerabilities in MMZ Transport Gear	3
	410	DoS Vulnerabilities in Internal Transport Gear	4

Firewalls make your network hard & crunchy on the outside, soft & chewy on the inside. –paraphrased from RFC1636 (circa 1994)

Why do so few of us contribute to the Problem Queue?

Five (5) people have contributed Problems in 2012

90+% from Stuart

I'm overwhelmed, I don't have time

Why bother? We don't' have time to fix these things

I don't understand any more what a Problem is, so I just keep my own list

I have too much to do already; if I submit something, I'll have to fix it

We don't want to admit to them – it's embarrassing

How do we struggle with the definition of Problem?

Examples

Servers Run Unsupported Operating Systems

We don't own the applications and therefore cannot drive upgrades

SCCA VPN Tunnels Vulnerable to Disruption

We've been transitioning services to the SCCA for years, makes it easy to ignore

Monitoring System does not detect Multiple Simultaneous UPS Failures

Facilities has asked for this enhancement but have no advocate inside CIT, so we ignore it

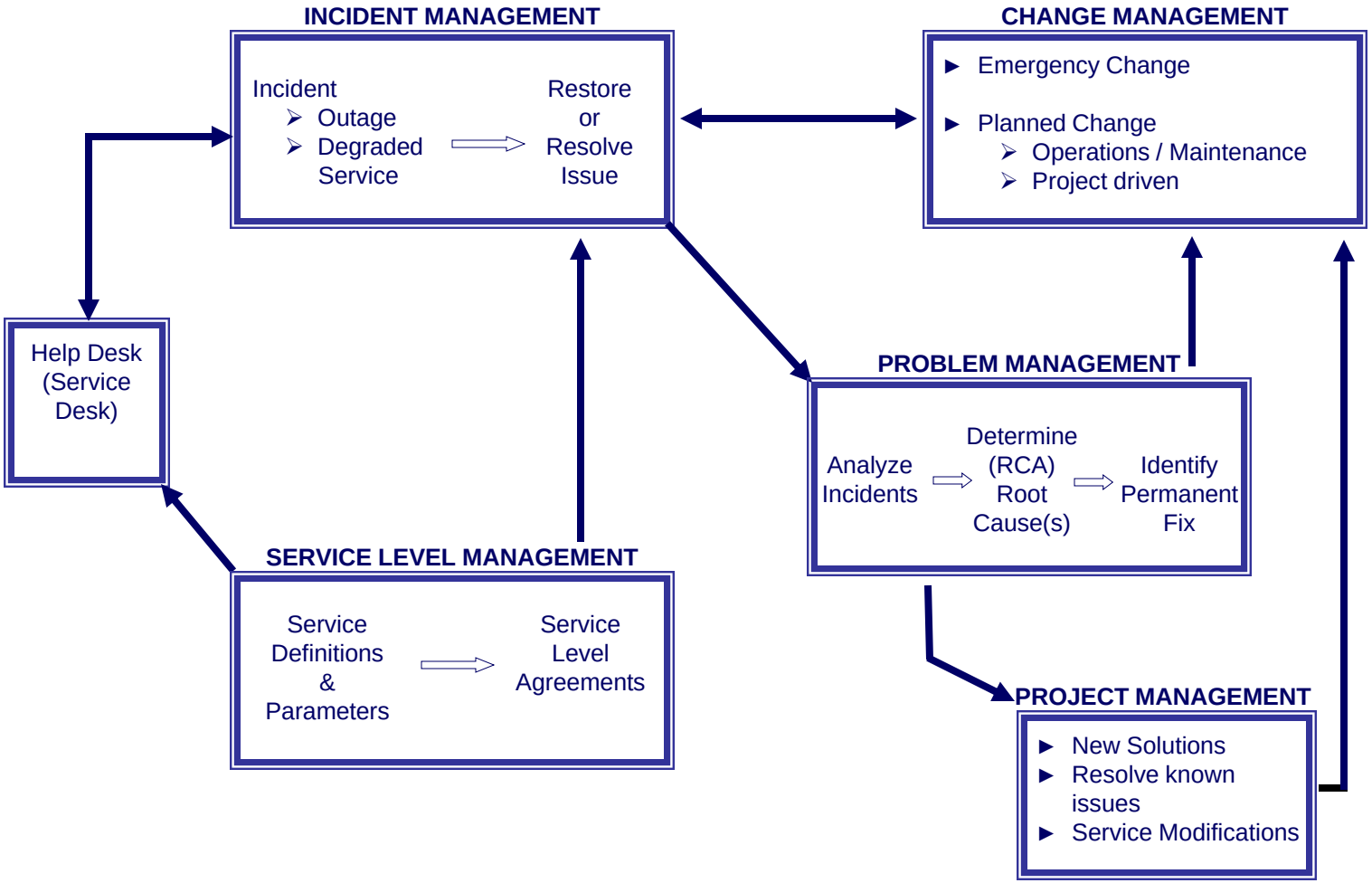
Philosophy

If it could cause an Incident, then it is a Problem

versus

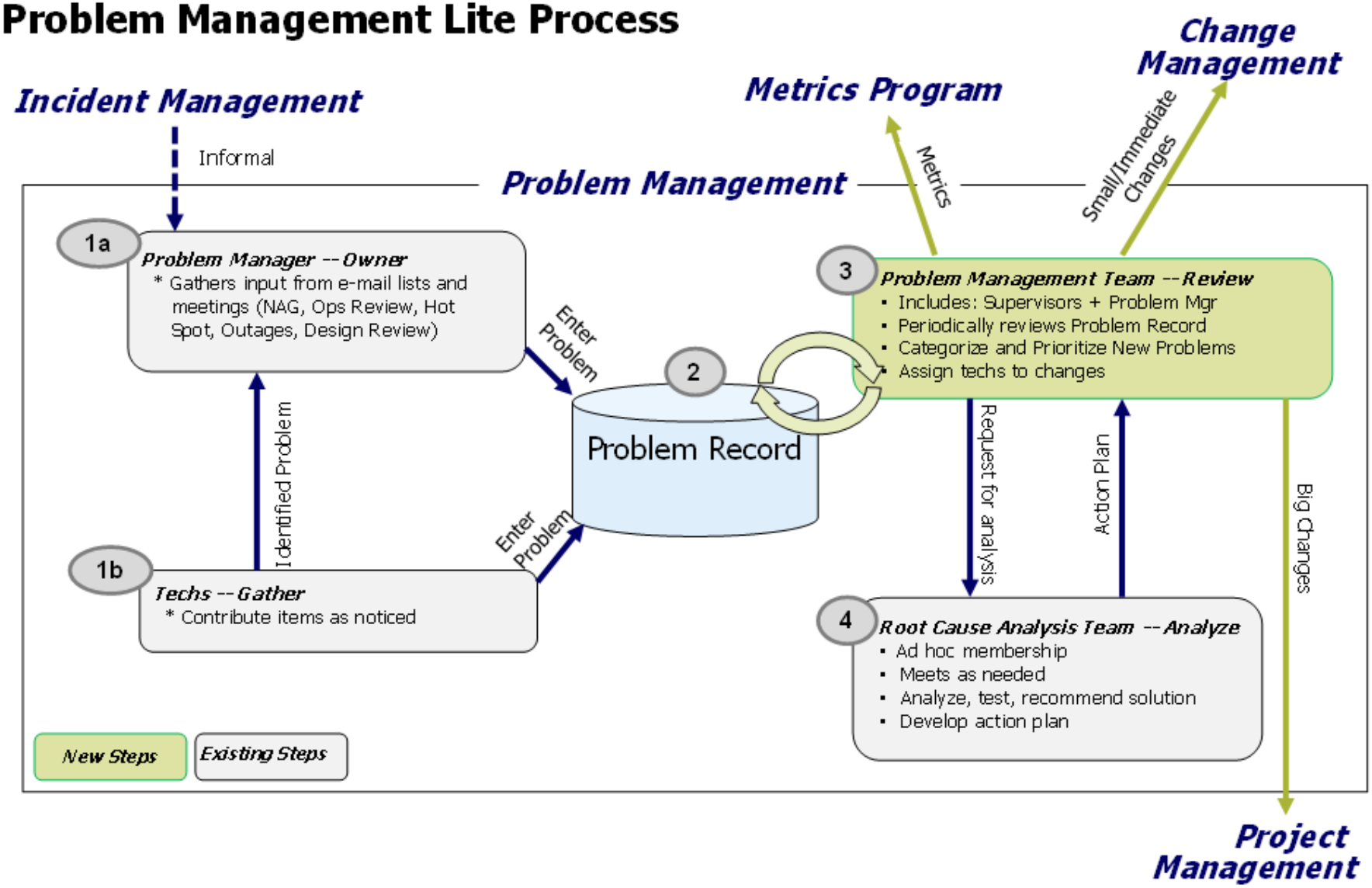
If it has not yet caused an Incident, then it cannot be a Problem

Service Support Framework
(Draft 12-07-10)



jhunter

Problem Management Lite Process



Priority Matrix

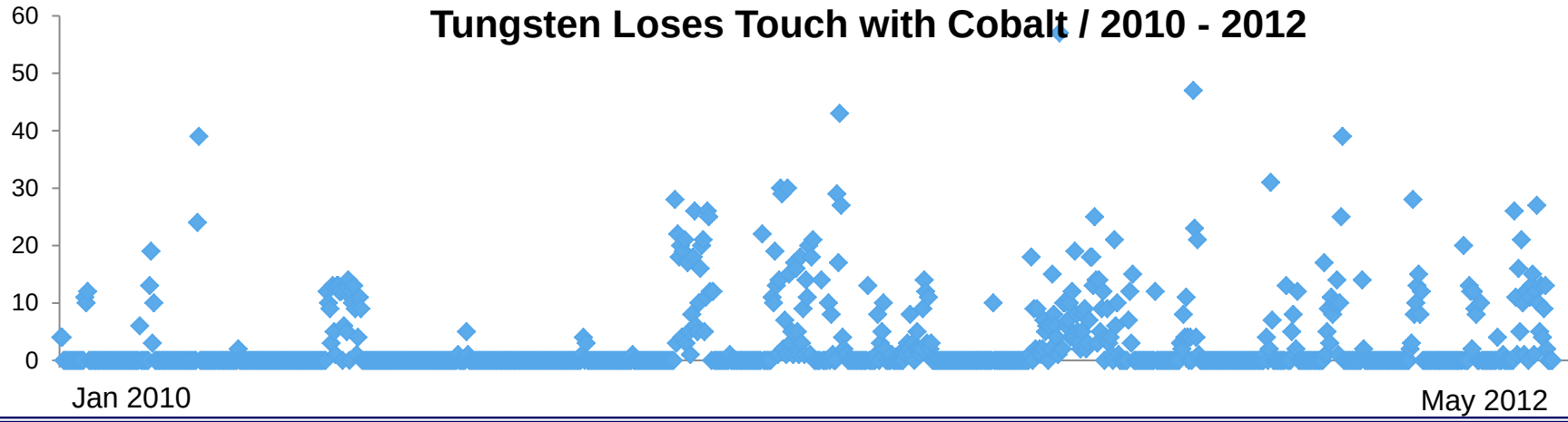
Priority is the relative urgency of the problem , calculated using the look-up table below

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

Tungsten Loses Touch With Cobalt

```
Jul  4 13:50:37 tungsten-a-svif1 [tungsten-a: cf.fsm.takeoverOfPartnerDisabled:notice]: Cluster monitor: takeover of tungsten-b
          disabled (status of backup mailbox is uncertain)
Jul  4 13:50:44 tungsten-b-svif1 [tungsten-b: cf.fsm.takeoverOfPartnerDisabled:notice]: Cluster monitor: takeover of tungsten-a
          disabled (status of backup mailbox is uncertain)
Jul  4 14:15:44 tungsten-b-svif1 [tungsten-b: cf.fsm.takeoverOfPartnerDisabled:notice]: Cluster monitor: takeover of tungsten-a
          disabled (status of backup mailbox is uncertain)
Jul  4 14:15:43 tungsten-a-svif1 [tungsten-a: cf.fsm.takeoverOfPartnerDisabled:notice]: Cluster monitor: takeover of tungsten-b
          disabled (status of backup mailbox is uncertain)
Jul  4 14:16:23 tungsten-b-svif1 [tungsten-b: cf.fsm.takeoverOfPartnerDisabled:notice]: Cluster monitor: takeover of tungsten-a
          disabled (status of backup mailbox is uncertain)
Jul  4 14:16:29 tungsten-a-svif1 [tungsten-a: cf.fsm.takeoverOfPartnerDisabled:notice]: Cluster monitor: takeover of tungsten-b
          disabled (status of backup mailbox is uncertain)
…
Jul 4 17:26:35 tungsten-a: isp2400_timeout_2: fci.device.quiesce:info: Adapter 0a encountered a command timeout on Disk device 0a.0
          (0x03000000) LUN 190 cdb 0x28:59c9b35c:00c0 retry: 0
Jul 4 17:26:35 tungsten-a: isp2400_timeout_2: scsi.path.excessiveErrors:err: Adapter 0a Excessive errors encountered by adapter disk
          device 0a.0
Jul 4 17:26:39 tungsten-a: isp2400_timeout_0: fci.device.quiesce:info: Adapter 0c encountered a command timeout on Disk device 0c.0
          (0x01000000) LUN 91 cdb 0x28:6948e865:0009 retry: 0
Jul 4 17:27:10 tungsten-a: isp2400_timeout_2: fci.device.timeout:err: HBA 0a encountered a device timeout on Disk 0a.0 (0x03000000)
          104 0x28:786b532e:0009 0
Jul 4 17:27:11 tungsten-a: isp2400_intrd: scsi.cmd.abortedByHost::err Disk device 0a.0L104: Command aborted by host adapter: HA
          status 0x4: cdb 0x28:786b532e:0009
…
```



Why is this group performing analysis?
Are you confident that you can do this better than your techs?
How about if we return to process engineering?

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE