Production Improvement Review (PIR) @ MS

Martin Check
mcheck@Microsoft.com
Microsoft Azure SRE
@mchecksre #AzureSRE

Azure SRE

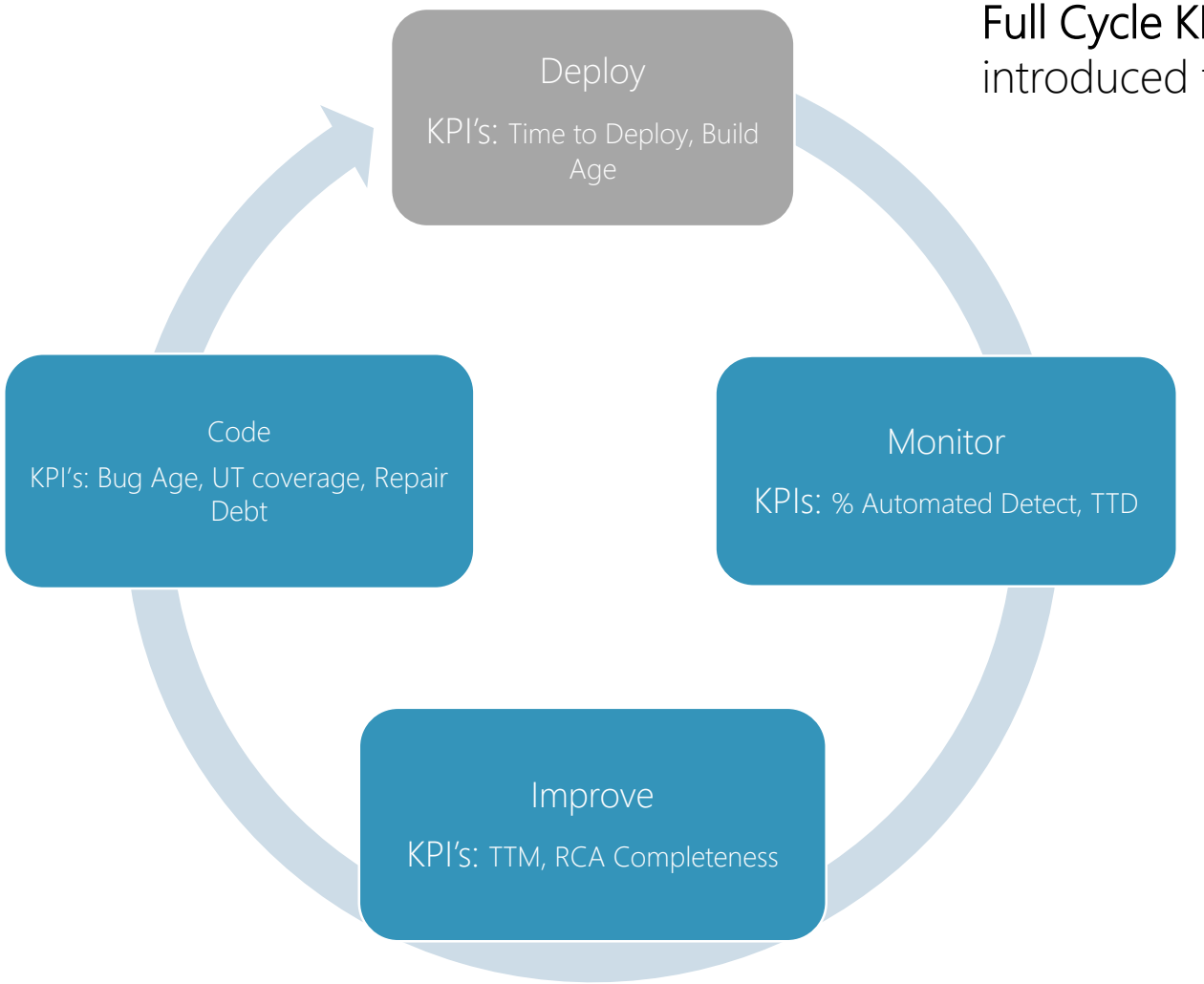
Why Production Improvement Review?

- Services with high incident rates
- Varying Service Maturity
- Inconsistent response models
- Competing improvement priorities

What is the Production Improvement Review?

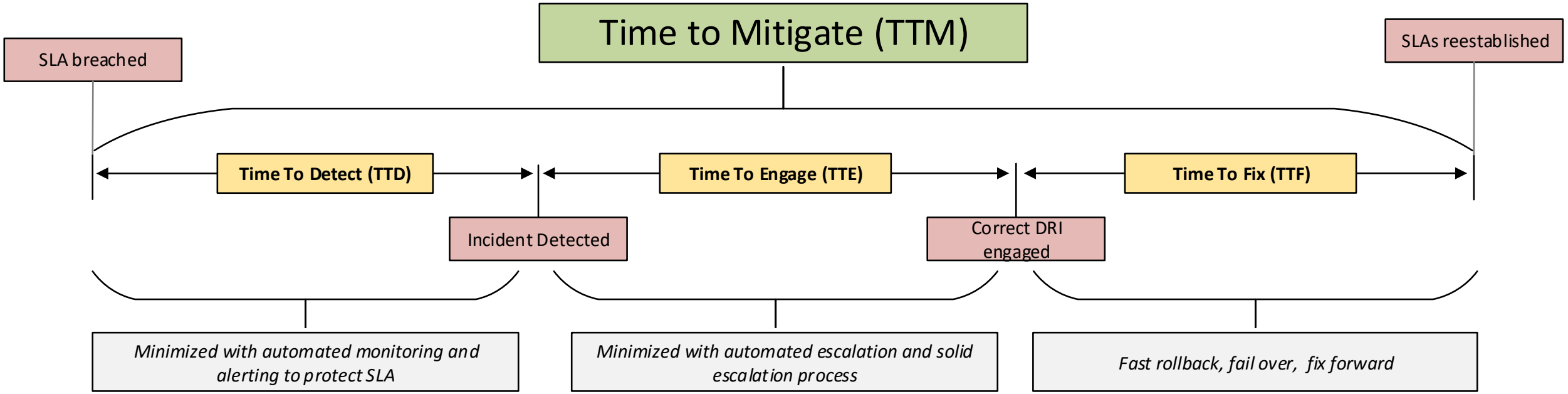
- Engineering focused meeting (not execs)
- Metrics Driven
- Use agreed upon, common KPI's aligned to top level objectives
  - Eliminate human touches
  - Availability / reliability
  - Velocity
- Action-Oriented
- Continuous Cadence

The Virtuous Cycle

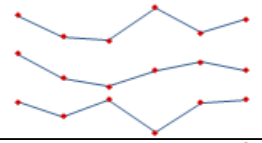
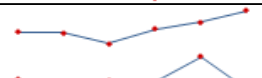
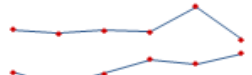
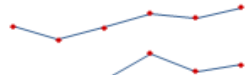
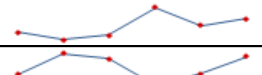
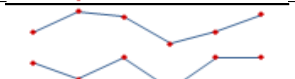
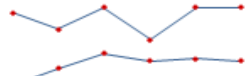


Full Cycle KPI: Time from bug introduced to fix rolled out worldwide

Defining Consistent Incident Response KPI's



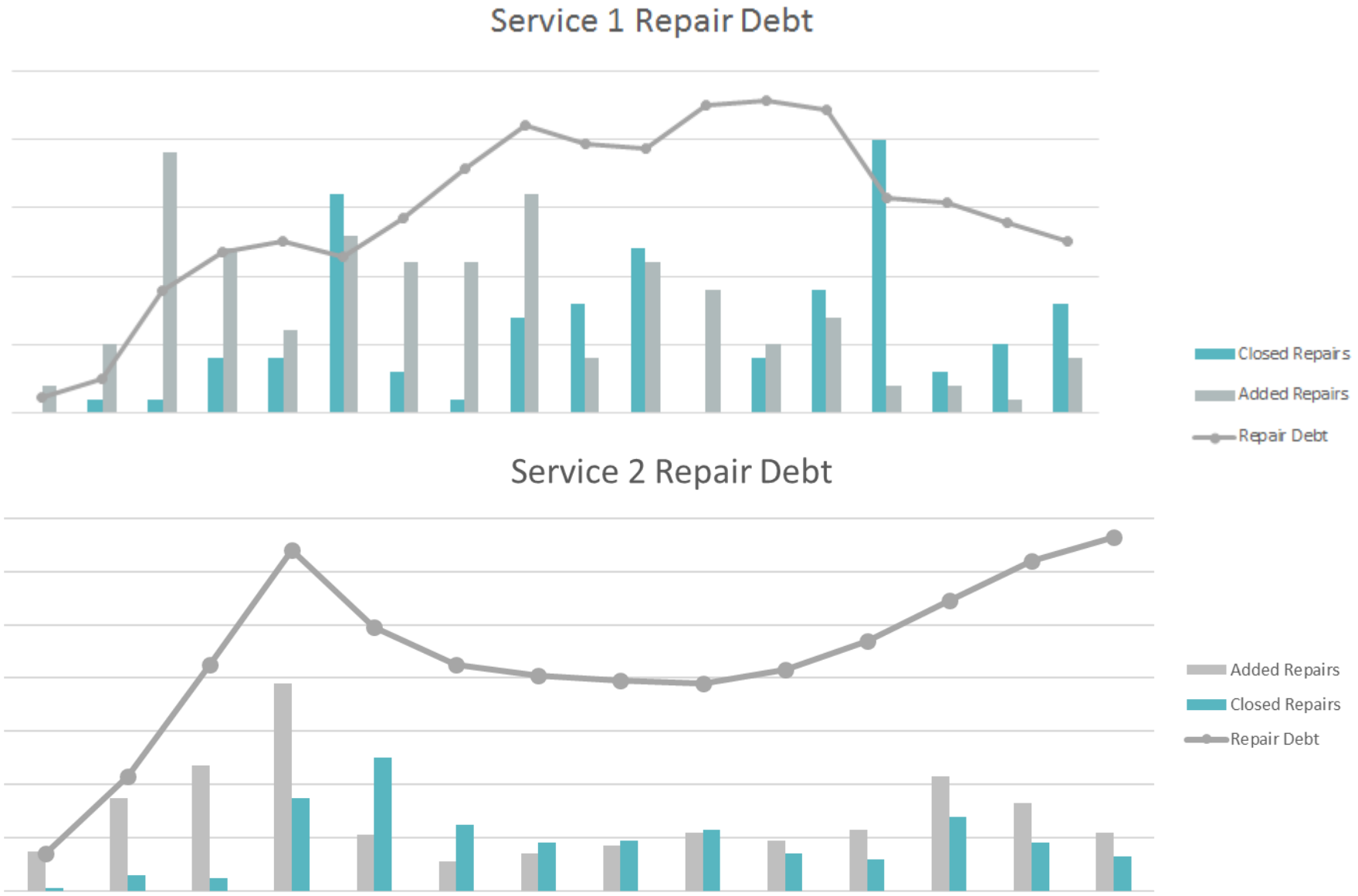
PIR Metrics Dashboard

	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Trend	Goal
Σ Incidents	XX	XX	XX	XX	XX	**XX.XXX%**		
Σ Major Incidents	X	X	X	X	X	**X**		
SLO	XX.XXX%	XX.XXX%	XX.XXX%	XX.XXX%	XX.XXX%	**XX.XXX%**		XX.XX%
TTD @ XX%ile	XX	XX	XX	XX	XX	**XX**		<X min
TTE @ XX%ile	XX	XX	XX	XX	XX	XX		<XX min
TTF @ XX%ile	XX	XXX	XX	XXX	XX	**XX**		<XX min
TTM @ XX%ile	XX	XXX	XX	XXX	XX	**XX**		<XX min
% Outages autodetected	XX%	XX%	XX%	XX%	XX%	**XX%**		XX%
# DRIs engaged per Bridge	XX	X	XX	X	XX	**XX**		X
DRI Hops	X	X	X	X	X	**X**		X

Top Incidents	Cause	TTD (mins)	TTM (mins)	Repair Items	Impact (reported)	Impact (Actual)
Incident in North Europe due to Code Bug	Code Bug	XX	XX	1	2	XXX Accounts Impacted
Network Incident due to Configuration	Config	X	XX	4	0	X,XXX Accounts impacted

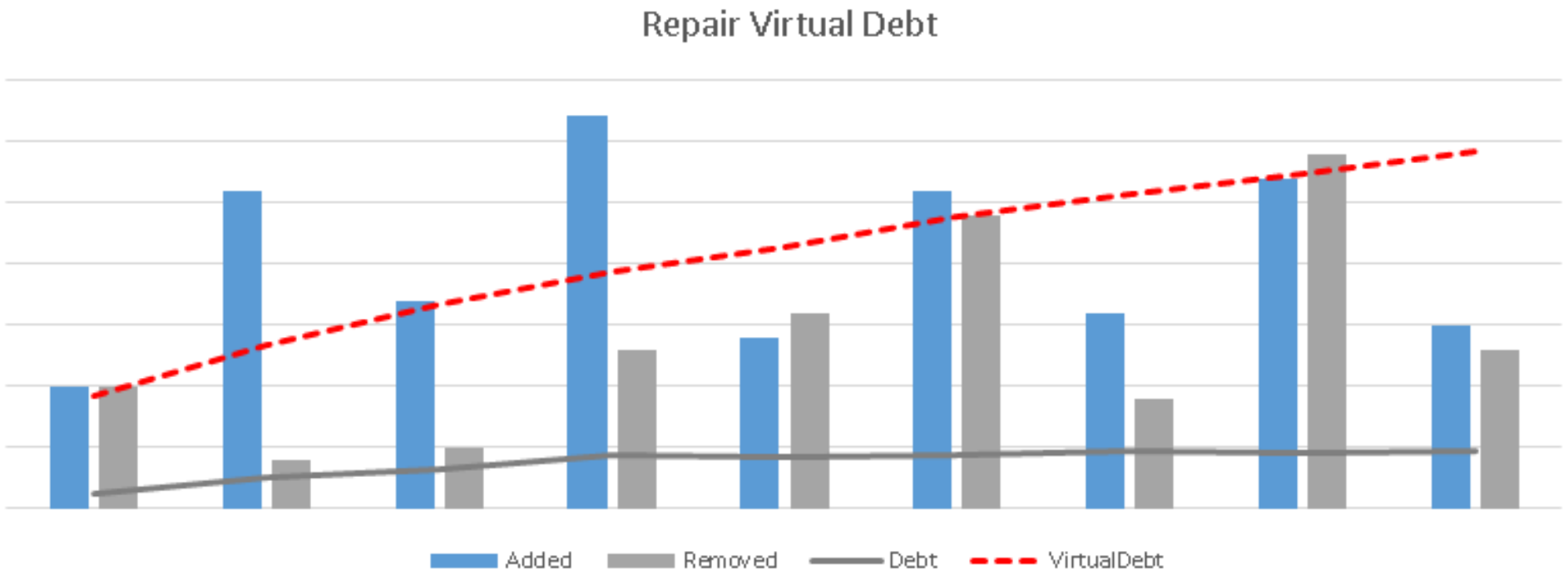
Deployment	95% of clusters		100% of clusters	
	Build Age	Build Age Trend	Build Age	Build Age Trend
Service A	XX		XXX	
Service B	XX		XX	

PIR Repair Debt

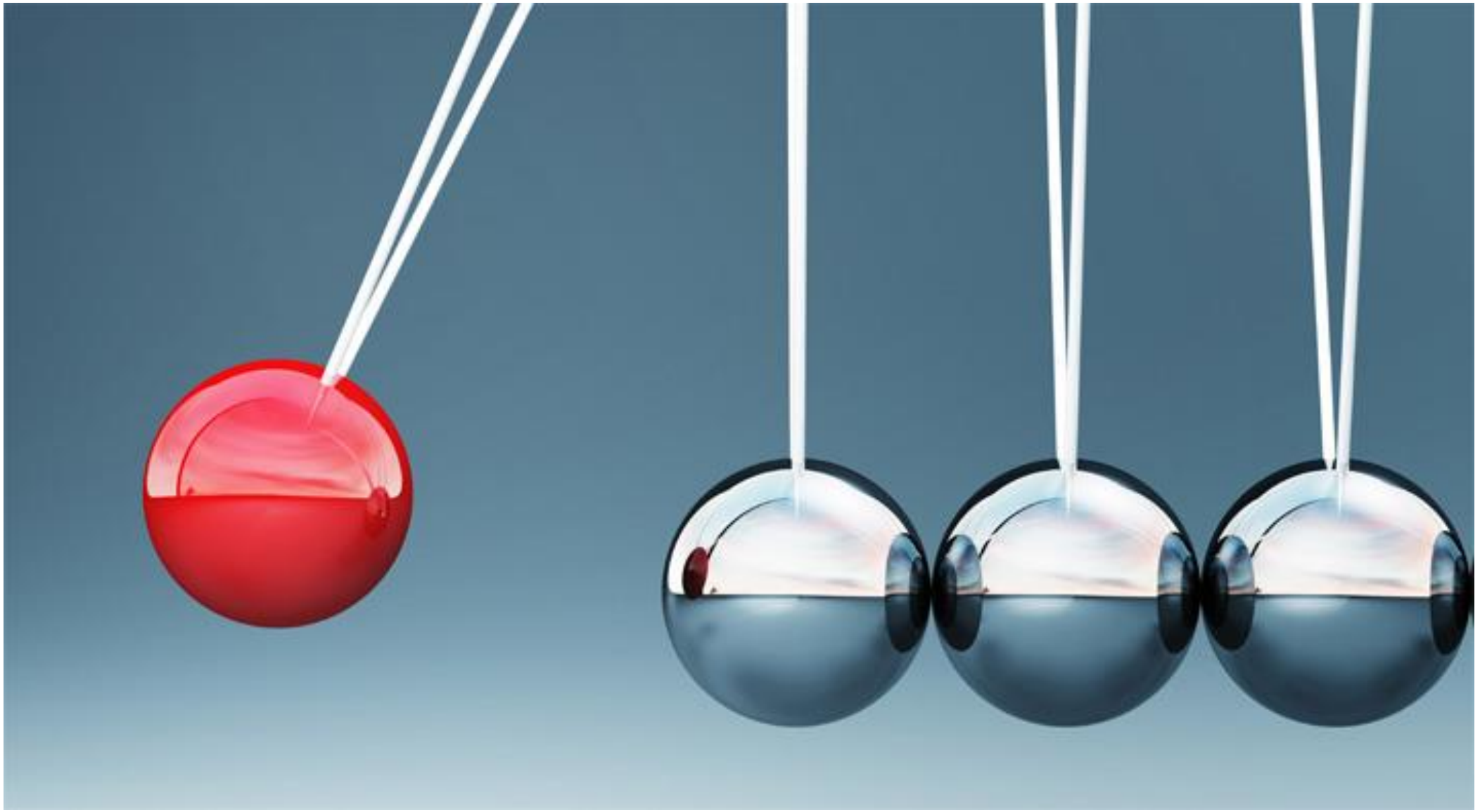


Repair Virtual Debt

Service Team	# incidents	# incidents with incomplete RCA	TTD misses	TTD repair items missing
Service 1	13	2	12	3
Service 2	8	1	6	2
Service 3	17	7	14	7
Service …				



Cadences



Real-time per service dashboards

Owning Service Group	Owning Service	#Incidents	RCA Completed	Repair Added
Service Group 1 (XX)	Service A	XX	XXX% (XX/XX)	XXX% (XX/XX)
	Service B	XX	XX% (XX/XX)	XX% (XX/XX)
	Service C	XX	XX% (XX/XX)	XXX% (XX/XX)
	Service D	X	XX% (X/X)	XX% (X/X)
	Service E	X	XXX% (X/X)	XXX% (X/X)
	Service F	X	XX% (X/X)	XXX% (X/X)
	Service G	X	XX% (X/X)	XX% (X/X)
Service Group 2 (XX)	Service H	XX	XX% (XX/XX)	XX% (XX/XX)
Service Group 3 (XX)	Service I	XX	XX% (XX/XX)	XX% (XX/XX)
Service Group …(XX)	Service J	XX	X% (X/XX)	XX% (X/XX)
	Service K	XX	X% (X/XX)	XX% (X/XX)
	Service L	X	XX% (X/X)	XXX% (X/X)

Insights (Expected and Unexpected)

- If you find a monitoring gap then try to fix it before you send everyone home
  - Incidents not detected by monitoring take 10-15X longer to mitigate
- Don’t fall into debug camp
- Right-size your repair
  - If the ultimate repair is large, try to find a quick version
  - But don’t forget it! – maintain an operational backlog to keep track of big items
- Taxonomy matters – human error is never human error
- Look out for Super DRI’s

Questions?

Martin Check

mcheck@Microsoft.com

@mchecksre #AzureSRE