Lessons Learned from Other Industries

Written by Jennifer Petoff
Edited by Betsy Beyer

A deep dive into SRE culture and practices at Google naturally leads to the question of how other industries manage their businesses for reliability. Compiling this book on Google SRE created an opportunity to speak to a number of Google's engineers about their previous work experiences in a variety of other high-reliability fields in order to address the following comparative questions:

- Are the principles used in Site Reliability Engineering also important outside of Google, or do other industries tackle the requirements of high reliability in markedly different ways?
- If other industries also adhere to SRE principles, how are the principles manifested?
- What are the similarities and differences in the implementation of these principles across industries?
- What factors drive similarities and differences in implementation?
- What can Google and the tech industry learn from these comparisons?

A number of principles fundamental to Site Reliability Engineering at Google are discussed throughout this text. To simplify our comparison of best practices in other industries, we distilled these concepts into four key themes:

- Preparedness and Disaster Testing
- Postmortem Culture
- Automation and Reduced Operational Overhead
- Structured and Rational Decision Making

This chapter introduces the industries that we profiled and the industry veterans we interviewed. We define key SRE themes, discuss how these themes are implemented at Google, and give examples of how these principles reveal themselves in other industries for comparative purposes. We conclude with some insights and discussion on the patterns and anti-patterns we discovered.

Meet Our Industry Veterans

Peter Dahl is a Principal Engineer at Google. Previously, he worked as a defense contractor on several high-reliability systems including many airborne and wheeled vehicle GPS and inertial guidance systems. Consequences of a lapse in reliability in such systems include vehicle malfunction or loss, and the financial consequences associated with that failure.

Mike Doherty is a Site Reliability Engineer at Google. He worked as a lifeguard and lifeguard trainer for a decade in Canada. Reliability is absolutely essential by nature in this field, because lives are on the line every day.

Erik Gross is currently a software engineer at Google. Before joining the company, he spent seven years designing algorithms and code for the lasers and systems used to perform refractive eye surgery (e.g., LASIK). This is a high-stakes, high-reliability field, in which many lessons relevant to reliability in the face of government regulations and human risk were learned as the technology received FDA approval, gradually improved, and finally became ubiquitous.

Gus Hartmann and **Kevin Greer** have experience in the telecommunications industry, including maintaining the E911 emergency response system.[154] Kevin is currently a software engineer on the Google Chrome team and Gus is a systems engineer for Google's Corporate Engineering team. User expectations of the telecom industry demand high reliability. Implications of a lapse of service range from user inconvenience due to a system outage to fatalities if E911 goes down.

Ron Heiby is a Technical Program Manager for Site Reliability Engineering at Google. Ron has experience in development for cell phones, medical devices, and the automotive industry. In some cases he worked on interface components of these industries (for example, on a device to allow EKG readings[155] in ambulances to be transmitted over the digital wireless phone network). In these industries, the impact of a reliability issue can range from harm to the business incurred by equipment recalls to indirectly impacting life and health (e.g., people not getting the medical attention they need if the EKG cannot communicate with the hospital).

Adrian Hilton is a Launch Coordination Engineer at Google. Previously, he worked on UK and USA military aircraft, naval avionics and aircraft stores management systems, and UK railway signaling systems. Reliability is critical in this space because impact of incidents ranges from multimillion-dollar loss of equipment to injuries and fatalities.

Eddie Kennedy is a project manager for the Global Customer Experience team at Google and a mechanical engineer by training. Eddie spent six years working as a Six Sigma Black Belt process engineer in a manufacturing facility that makes synthetic diamonds. This industry is characterized by a relentless focus on safety, because the extremes of temperature and pressure demands of the process pose a high level of danger to workers on a daily basis.

John Li is currently a Site Reliability Engineer at Google. John previously worked as a systems administrator and software developer at a proprietary trading company in the finance industry. Reliability issues in the financial sector are taken quite seriously because they can lead to serious fiscal consequences.

Dan Sheridan is a Site Reliability Engineer at Google. Before joining the company, he worked as a safety consultant in the civil nuclear industry in the UK. Reliability is important in the nuclear industry because an incident can have serious repercussions: outages can incur millions a day in lost revenue, while risks to workers and those in the community are even

more dire, dictating zero tolerance for failure. Nuclear infrastructure is designed with a series of failsafes that halt operations before an incident of any such magnitude is reached.

Jeff Stevenson is currently a hardware operations manager at Google. He has past experience as a nuclear engineer in the US Navy on a submarine. Reliability stakes in the nuclear Navy are high—problems that arise in the case of incidents range from damaged equipment, to long-standing environmental impact, to potential loss of life.

Matthew Toia is a Site Reliability Manager focused on storage systems. Prior to Google, he worked on software development and deployment of air traffic control software systems. Effects from incidents in this industry range from inconveniences to passengers and airlines (e.g., delayed flights, diverted planes) to potential loss of life in the event of a crash. Defense in depth is a key strategy to avoiding catastrophic failures.

Now that you've met our experts and gained a high-level understanding of why reliability is important in their respective former fields, we'll delve into the four key themes of reliability.

Preparedness and Disaster Testing

"Hope is not a strategy." This rallying cry of the SRE team at Google sums up what we mean by preparedness and disaster testing. The SRE culture is forever vigilant and constantly questioning: What could go wrong? What action can we take to address those issues before they lead to an outage or data loss? Our annual Disaster and Recovery Testing (DiRT) drills seek to address these questions head-on [Kri12]. In DiRT exercises, SREs push production systems to the limit and inflict actual outages in order to:

- Ensure that systems react the way we think they will
- Determine unexpected weaknesses
- Figure out ways to make the systems more robust in order to prevent uncontrolled outages

Several strategies for testing disaster readiness and ensuring preparedness in other industries emerged from our conversations. Strategies included the following:

- Relentless organizational focus on safety
- Attention to detail
- Swing capacity
- Simulations and live drills
- Training and certification
- Obsessive focus on detailed requirements gathering and design
- Defense in depth

Relentless Organizational Focus on Safety

This principle is particularly important in an industrial engineering context. According to Eddie Kennedy, who worked on a manufacturing floor where workers faced safety hazards, "every management meeting started with a discussion of safety." The manufacturing industry prepares itself for the unexpected by establishing highly defined processes that are strictly followed at every level of the organization. It is critical that all employees take safety seriously, and that workers feel empowered to speak up if and when anything seems amiss. In the case of nuclear power, military aircraft, and railway signaling industries, safety standards for software are well detailed (e.g., UK Defence Standard 00-56, IEC 61508, IEC513, US DO-178B/C, and DO-254) and levels of reliability for such systems are clearly identified (e.g., Safety Integrity Level (SIL) 1–4),[156] with the aim of specifying acceptable approaches to delivering a product.

Attention to Detail

From his time spent in the US Navy, Jeff Stevenson recalls an acute awareness of how a lack of diligence in executing small tasks (for example, lube oil maintenance) could lead to major submarine failure. A very small oversight or mistake can have big effects. Systems are highly interconnected, so an accident in one area can impact multiple related components. The nuclear Navy focuses on routine maintenance to ensure that small issues don't snowball.

Swing Capacity

System utilization in the telecom industry can be highly unpredictable. Absolute capacity can be strained by unforeseeable events such as natural disasters, as well as large, predictable events like the Olympics. According to Gus Hartmann, the industry deals with these incidents by deploying swing capacity in the form of a SOW (switch on wheels), a mobile telco office. This excess capacity can be rolled out in an emergency or in anticipation of a known event that is likely to overload the system. Capacity issues veer into the unexpected in matters unrelated to absolute capacity, as well. For example, when a celebrity's private phone number was leaked in 2005 and thousands of fans simultaneously attempted to call her, the telecom system exhibited symptoms similar to a DDoS or massive routing error.

Simulations and Live Drills

Google's Disaster Recovery tests have a lot in common with the simulations and live drills that are a key focus of many of the established industries we researched. The potential consequences of a system outage determine whether using a simulation or a live drill is appropriate. For example, Matthew Toia points out that the aviation industry can't perform a live test "in production" without putting equipment and passengers at risk. Instead, they employ extremely realistic simulators with live data feeds, in which the control rooms and equipment are modeled down to the tiniest details to ensure a realistic experience without putting real people at risk. Gus Hartmann reports that the telecom industry typically focuses on live drills centered on surviving hurricanes and other weather emergencies. Such modeling led to the creation of weatherproof facilities with generators inside the building capable of outlasting a storm.

The US nuclear Navy uses a mixture of "what if" thought exercises and live drills. According to Jeff Stevenson, the live drills involve "actually breaking real stuff but with control parameters. Live drills are carried out religiously, every week, two to three days per week."For the nuclear Navy, thought exercises are useful, but not sufficient to prepare for actual incidents. Responses must be practiced so they are not forgotten.

According to Mike Doherty, lifeguards face disaster testing exercises more akin to a "mystery shopper" experience. Typically, a facility manager works with a child or an incognito lifeguard

in training to stage a mock drowning. These scenarios are conducted to be as realistic as possible so that lifeguards aren't able to differentiate between real and staged emergencies.

Training and Certification

Our interviews suggest that training and certification are particularly important when lives are at stake. For example, Mike Doherty described how lifeguards complete a rigorous training certification, in addition to a periodic recertification process. Courses include fitness components (e.g., a lifeguard must be able to hold someone heavier than themselves with shoulders out of the water), technical components like first aid and CPR, and operational elements (e.g., if a lifeguard enters the water, how do other team members respond?). Every facility also has site-specific training, because lifeguarding in a pool is markedly different from lifeguarding on a lakeside beach or on the ocean.

Focus on Detailed Requirements Gathering and Design

Some of the engineers we interviewed discussed the importance of detailed requirements gathering and design docs. This practice was particularly important when working with medical devices. In many of these cases, actual use or maintenance of the equipment doesn't fall within the purview of product designers. Thus, usage and maintenance requirements must be gathered from other sources.

For example, according to Erik Gross, laser eye surgery machines are designed to be as foolproof as possible. Thus, soliciting requirements from the surgeons who actually use these machines and the technicians responsible for maintaining them is particularly important. In another example, former defense contractor Peter Dahl described a very detailed design culture in which creating a new defense system commonly entailed an entire year of design, followed by just three weeks of writing the code to actualize the design. Both of these examples are markedly different from Google's launch and iterate culture, which promotes a much faster rate of change at a calculated risk. Other industries (e.g., the medical industry and the military, as previously discussed) have very different pressures, risk appetites, and requirements, and their processes are very much informed by these circumstances.

Defense in Depth and Breadth

In the nuclear power industry, defense in depth is a key element to preparedness [IAEA12]. Nuclear reactors feature redundancy on all systems and implement a design methodology that mandates fallback systems behind primary systems in case of failure. The system is designed with multiple layers of protection, including a final physical barrier to radioactive release around the plant itself. Defense in depth is particularly important in the nuclear industry due to the zero tolerance for failures and incidents.

Postmortem Culture

Corrective and preventative action (CAPA)[157] is a well-known concept for improving reliability that focuses on the systematic investigation of root causes of identified issues or risks in order to prevent recurrence. This principle is embodied by SRE's strong culture of blameless postmortems. When something goes wrong (and given the scale, complexity, and rapid rate of change at Google, something inevitably *will* go wrong), it's important to evaluate all of the following:

- What happened
- The effectiveness of the response
- What we would do differently next time
- What actions will be taken to make sure a particular incident doesn't happen again

This exercise is undertaken without pointing fingers at any individual. Instead of assigning blame, it is far more important to figure out *what* went wrong, and how, as an organization, we will rally to ensure it doesn't happen again. Dwelling on *who* might have caused the outage is counterproductive. Postmortems are conducted after incidents and published across SRE teams so that all can benefit from the lessons learned.

Our interviews uncovered that many industries perform a version of the postmortem (although many do not use this specific moniker, for obvious reasons). The *motivation* behind these exercises appears to be the main differentiator among industry practices.

Many industries are heavily regulated and are held accountable by specific government authorities when something goes wrong. Such regulation is especially ingrained when the stakes of failure are high (e.g., lives are at stake). Relevant government agencies include the FCC (telecommunications), FAA (aviation), OSHA (the manufacturing and chemical industries), FDA (medical devices), and the various National Competent Authorities in the EU. [158] The nuclear power and transportation industries are also heavily regulated.

Safety considerations are another motivating factor behind postmortems. In the manufacturing and chemical industries, the risk of injury or death is ever-present due to the nature of the conditions required to produce the final product (high temperature, pressure, toxicity, and corrosivity, to name a few). For example, Alcoa features a noteworthy safety culture. Former CEO Paul O'Neill required staff to notify him within 24 hours of any injury that lost a worker day. He even distributed his home phone number to workers on the factory floor so that they could personally alert him to safety concerns.[159]

The stakes are so high in the manufacturing and chemical industries that even "near misses"—when a given event could have caused serious harm, but did not—are carefully scrutinized. These scenarios function as a type of preemptive postmortem. According to VM Brasseur in a talk given at YAPC NA 2015, "There are multiple near misses in just about every disaster and business crisis, and typically they're ignored at the time they occur. Latent error, plus an enabling condition, equals things not working quite the way you planned" [Bra15]. Near misses are effectively disasters waiting to happen. For example, scenarios in which a worker doesn't follow the standard operating procedure, an employee jumps out of the way at the last second to avoid a splash, or a spill on the staircase isn't cleaned up, all represent near misses and opportunities to learn and improve. Next time, the employee and the company might not be so lucky. The United Kingdom's CHIRP (Confidential Reporting Programme for Aviation and Maritime) seeks to raise awareness about such incidents across the industry by providing a central reporting point where aviation and maritime personnel can report near misses confidentially. Reports and analyses of these near misses are then published in periodic newsletters.

Lifeguarding has a deeply embedded culture of post-incident analysis and action planning. Mike Doherty quips, "If a lifeguard's feet go in the water, there will be paperwork!" A detailed write-up is required after any incident at the pool or on the beach. In the case of serious incidents, the team collectively examines the incident end to end, discussing what went right and what went wrong. Operational changes are then made based on these findings, and training is often scheduled to help people build confidence around their ability to handle a similar incident in the future. In cases of particularly shocking or traumatic incidents, a counselor is brought on site to help staff cope with the psychological aftermath. The lifeguards may have been well prepared for what happened in practice, but might *feel* like they haven't done an adequate job. Similar to Google, lifeguarding embraces a culture of blameless incident analysis. Incidents are chaotic, and many factors contribute to any given incident. In this field, it's not helpful to place blame on a single individual.

Automating Away Repetitive Work and Operational Overhead

At their core, Google's Site Reliability Engineers are software engineers with a low tolerance for repetitive reactive work. It is strongly ingrained in our culture to avoid repeating an operation that doesn't add value to a service. If a task can be automated away, why would you run a system on repetitive work that is of low value? Automation lowers operational overhead and frees up time for our engineers to proactively assess and improve the services they support.

The industries that we surveyed were mixed in terms of if, how, and why they embraced automation. Certain industries trusted humans more than machines. During the tenure of our industry veteran, the US nuclear Navy eschewed automation in favor of a series of interlocks and administrative procedures. For example, according to Jeff Stevenson, operating a valve required an operator, a supervisor, and a crew member on the phone with the engineering watch officer tasked with monitoring the response to the action taken. These operations were very manual due to concern that an automated system might not spot a problem that a human would definitely notice. Operations on a submarine are ruled by a trusted human decision chain—a *series* of people, rather than one individual. The nuclear Navy was also concerned that automation and computers move so rapidly that they are all too capable of

committing a large, irreparable mistake. When you are dealing with nuclear reactors, a slow and steady methodical approach is more important than accomplishing a task quickly.

According to John Li, the proprietary trading industry has become increasingly cautious in its application of automation in recent years. Experience has shown that incorrectly configured automation can inflict significant damage and incur a great deal of financial loss in a very short period of time. For example, in 2012 Knight Capital Group encountered a "software glitch" that led to a loss of $440M in just a few hours.[160] Similarly, in 2010 the US stock market experienced a Flash Crash that was ultimately blamed on a rogue trader attempting to manipulate the market with automated means. While the market was quick to recover, the Flash Crash resulted in a loss on the magnitude of trillions of dollars in just *30 minutes*.[161] Computers can execute tasks very quickly, and speed can be a negative if these tasks are configured incorrectly.

In contrast, some companies embrace automation precisely *because* computers act more quickly than people. According to Eddie Kennedy, efficiency and monetary savings are key in the manufacturing industry, and automation provides a means to accomplish tasks more efficiently and cost-effectively. Furthermore, automation is generally more reliable and repeatable than work conducted manually by humans, which means that it produces higher-quality standards and tighter tolerances. Dan Sheridan discussed automation as deployed in the UK nuclear industry. Here, a rule of thumb dictates that if a plant is required to respond to a given situation in less than 30 minutes, that response must be automated.

In Matt Toia's experience, the aviation industry applies automation selectively. For example, operational failover is performed automatically, but when it comes to certain other tasks, the industry trusts automation only when it's verified by a human. While the industry employs a good deal of automatic monitoring, actual air-traffic–control-system implementations must be manually inspected by humans.

According to Erik Gross, automation has been quite effective in reducing user error in laser eye surgery. Before LASIK surgery is performed, the doctor measures the patient using a refractive eye test. Originally, the doctor would type in the numbers and press a button, and the laser would go to work correcting the patient's vision. However, data entry errors could be a big issue. This process also entailed the possibility of mixing up patient data or jumbling numbers for the left and right eye.

Automation now greatly lessens the chance that humans make a mistake that impacts someone's vision. A computerized sanity check of manually entered data was the first major automated improvement: if a human operator inputs measurements outside an expected range, automation promptly and prominently flags this case as unusual. Other automated improvements followed this development: now the iris is photographed during the preliminary refractive eye test. When it's time to perform the surgery, the iris of the patient is automatically matched to the iris in the photo, thus eliminating the possibility of mixing up patient data. When this automated solution was implemented, an entire class of medical errors disappeared.

Structured and Rational Decision Making

At Google in general, and in Site Reliability Engineering in particular, data is critical. The team aspires to make decisions in a structured and rational way by ensuring that:

- The basis for the decision is agreed upon advance, rather than justified ex post facto
- The inputs to the decision are clear
- Any assumptions are explicitly stated
- Data-driven decisions win over decisions based on feelings, hunches, or the opinion of the most senior employee in the room

Google SRE operates under the baseline assumption that everyone on the team:

- Has the best interests of a service's users at heart
- Can figure out how to proceed based on the data available

Decisions should be informed rather than prescriptive, and are made without deference to personal opinions—even that of the most-senior person in the room, who Eric Schmidt and

Jonathan Rosenberg dub the "HiPPO," for "Highest-Paid Person's Opinion" [Sch14].

Decision making in different industries varies widely. We learned that some industries use an approach of *if it ain't broke, don't fix it…ever*. Industries featuring systems whose design entailed much thought and effort are often characterized by a reluctance to change the underlying technology. For example, the telecom industry still uses long-distance switches that were implemented in the 1980s. Why do they rely on technology developed a few decades ago? These switches "are pretty much bulletproof and massively redundant," according to Gus Hartmann. As reported by Dan Sheridan, the nuclear industry is similarly slow to change. All decisions are underpinned by the thought: *if it works now, don't change it*.

Many industries heavily focus on playbooks and procedures rather than open-ended problem solving. Every humanly conceivable scenario is captured in a checklist or in "the binder." When something goes wrong, this resource is the authoritative source for how to react. This prescriptive approach works for industries that evolve and develop relatively slowly, because the scenarios of what could go wrong are not constantly evolving due to system updates or changes. This approach is also common in industries in which the skill level of the workers may be limited, and the best way to make sure that people will respond appropriately in an emergency is to provide a simple, clear set of instructions.

Other industries also take a clear, data-driven approach to decision making. In Eddie Kennedy's experience, research and manufacturing environments are characterized by a rigorous experimentation culture that relies heavily on formulating and testing hypotheses. These industries regularly conduct controlled experiments to make sure that a given change yields the expected result at a statistically significant level and that nothing unexpected occurs. Changes are only implemented when data yielded by the experiment supports the decision.

Finally, some industries, like proprietary trading, divide decision making to better manage risk. According to John Li, this industry features an enforcement team separate from the traders to ensure that undue risks aren't taken in pursuit of achieving a profit. The enforcement team is responsible for monitoring events on the floor and halting trading if events spin out of hand. If a system abnormality occurs, the enforcement team's first response is to shut down the system. As put by John Li, "If we aren't trading, we aren't losing money. We aren't making money either, but at least we aren't losing money." Only the

enforcement team can bring the system back up, despite how excruciating a delay might seem to traders who are missing a potentially profitable opportunity.

Conclusions

Many of the principles that are core to Site Reliability Engineering at Google are evident across a wide range of industries. The lessons already learned by well-established industries likely inspired some of the practices in use at Google today.

A main takeaway of our cross-industry survey was that in many parts of its software business, Google has a higher appetite for velocity than players in most other industries. The ability to move or change quickly must be weighed against the differing implications of a failure. In the nuclear, aviation, or medical industries, for example, people could be injured or even die in the event of an outage or failure. When the stakes are high, a conservative approach to achieving high reliability is warranted.

At Google, we constantly walk a tightrope between user expectations for high reliability versus a laser-sharp focus on rapid change and innovation. While Google is incredibly serious about reliability, we must adapt our approaches to our high rate of change. As discussed in earlier chapters, many of our software businesses such as Search make conscious decisions as to how reliable "reliable enough" really is.

Google has that flexibility in most of our software products and services, which operate in an environment in which lives are not directly at risk if something goes wrong. Therefore, we're able to use tools such as error budgets (Motivation for Error Budgets) as a means to "fund" a culture of innovation and calculated risk taking. In essence, Google has adapted known reliability principles that were in many cases developed and honed in other industries to create its own unique reliability culture, one that addresses a complicated equation that balances scale, complexity, and velocity with high reliability.

---

[154] E911 (Enhanced 911): Emergency response line in the US that leverages location data.

[155]Electrocardiogram readings: *https://en.wikipedia.org/wiki/Electrocardiography*.

[156]*https://en.wikipedia.org/wiki/Safety_integrity_level*

[157]*https://en.wikipedia.org/wiki/Corrective_and_preventive_action*

[158]*https://en.wikipedia.org/wiki/Competent_authority*

[159]*https://ehstoday.com/safety/nsc-2013-oneill-exemplifies-safety-leadership*.

[160]See "FACTS, Section B" for the discussion of Knight and Power Peg software in [Sec13].

[161]"Regulators blame computer algorithm for stock market 'flash crash'," Computerworld, *https://www.computerworld.com/article/2516076/financial-it/regulators-blame-computer-algorithm-for-stock-market—flash-crash-.html*.

PREVIOUS
Part V - Conclusions

NEXT
Chapter 34 - Conclusion