Seven Failure Scenarios:

What lessons can be learnt from recent high profile data centre failures?
Where are we relative to our Peer Industries?
Topics for Discussion

Failure Scenarios

• What failed?
• Why did it fail?
• Failure prevention
• Responsibility
• Lesson Learned

Failure Modelling & Analysis

• Understanding Failure (Swiss Cheese)
• Equipment Failures
• Failure Modes & Effects Analysis
• Universal Learning Curve
• Human Bathtub Curve
<table>
<thead>
<tr>
<th>Failure</th>
<th>Who</th>
<th>What</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Design</td>
<td>Consulting Engineer</td>
<td>Human</td>
</tr>
<tr>
<td>Systems Validation</td>
<td>Commissioning Engineer</td>
<td>Human</td>
</tr>
<tr>
<td>Equipment Design</td>
<td>OEM</td>
<td>Human</td>
</tr>
<tr>
<td>Equipment Failure</td>
<td>No one</td>
<td>Machine</td>
</tr>
<tr>
<td>Operator Error</td>
<td>Ops</td>
<td>Human</td>
</tr>
<tr>
<td>Installation Error</td>
<td>Contractor</td>
<td>Human</td>
</tr>
<tr>
<td>Maintenance Oversight</td>
<td>Ops</td>
<td>Human</td>
</tr>
<tr>
<td>Natural Disaster</td>
<td>Act of God</td>
<td>Various</td>
</tr>
</tbody>
</table>
Data centre failures are often the result of two or sometimes three simultaneous events.

Data centres are bespoke complex homo-technical systems.

Data centre failures are generally non-fatal, unlike other industries such as aviation.

This will probably change as human dependency on IT increases further.

Currently the data centre industry is unregulated.

Root cause investigation findings are normally secret and bound by NDA.

The data centre industry is NOT learning from its failures.

Industry is currently nowhere near the upper practical limit of reliability of 100,000 to 200,000 hours.
Engine No 2 explodes 4 minutes after take off

Fuel leak in the left mid fuel tank (the beast has 11 tanks, including in the horizontal stabilizer on the tail)

Fuel leak in the left inner fuel tank

Hole on the flap canoe/fairing that you could fit your upper body through

Aft gallery in the fuel system failed, preventing many fuel transfer functions

Fuel jettison had problems due to the problem above

Hole in the upper wing surface

Partial failure of leading edge slats

Partial failure of speed brakes/ground spoilers

Shrapnel damage to the flaps
QF 32 4 Nov 2010 – What Happened?

Figure 4: Close up of damage between at leading edge ribs 4 showing droop nose drive motor

Figure 5: Forward view of engine and subsequent damage for this item.
Figure 1: WU-1 Perforation at topskin Rib 12 location
Root Cause of the Failure

- Total loss of hydraulic fluid in the green system (2 x 5,000 PSI systems, green and yellow)
- Loss of one generator and associated systems
- Loss of brake anti-skid system
- Unable to shutdown adjacent Engine No 1 using normal method after landing due to major damage to systems
- Unable to shutdown adjacent engine using fire switch
- No fire protection was available for that engine after the explosion in Engine No2
- ECAM warnings about major fuel imbalance because of fuel leaks on left side, that were UNABLE to be fixed with cross-feeding
- Fuel trapped in trim tank (in the tail). Therefore, possible major C of G issue
Root Cause of the Failure

Australian Transport Safety Bureau indicated

- "Fatigue cracking" in a stub pipe within the engine
- resulted in oil leakage followed by an oil fire in the engine
- The fire led to the release of the intermediate pressure turbine (IPT) disc

Rolls Royce determined that the direct cause of the oil fire and resulting engine failure was a:

**Misaligned counter bore within a stub oil pipe leading to a fatigue fracture**

Plane lands 90 minutes later - No fatalities, No injuries

i.e. EQUIPMENT DESIGN – HUMAN ERROR
Skill, Experience, Training and Rigorous Operational Processes

- Captain was in the left seat, FO in the right, SO in the 2nd obs seat (right rear, also with his own Radio Management Panel, so he probably did most of the coordination with the ground), Capt Dave Evans in the 1st obs seat (middle). He is a Check & Training Captain who was training Harry Wubbin to be one also. Harry was in the 3rd obs seat (left rear).

- All 5 guys were FLAT OUT especially the FO who would have been processing complicated 'ECAM' messages and procedures that were seemingly never-ending!
(N – 1) Generators

- Hot summer day, utility power outage, data centre at full load 7.2MW
- Four 2.5 MW generators installed (N+1 configured)
- One generator fails to start (now N configured)
- Running on 3 generators
- 30 minutes later another generator fails (now N-1 configured)
- Now 5MW capacity supporting a 7.2MW load
- Remaining generators overload \(\approx\) 60 seconds
- Cooling plant has no power
- IT equipment begins shutting down (over-temperature)
- DC data centre runs on UPS for another 30mins (2N 15mins each side)
- Total data centre failure 30 minutes later
- Utility restored after 6 hours
- Data Centre fully restored after 8 hours later
- Worldwide enquiry launched
- Litigation and financial penalties
- Highly publicized therefore reputational damage
Lessons Learned from (N – 1) Generators

What Failed?

• Power and cooling leading to a complete data centre failure
• Massive international disruption to global internet traffic

Why did it fail?

• Lack of maintenance to seals in the pneumatic starter
• Generators were maintained but air system maintenance was excluded
• Causal factors: utility outage, generator pneumatic starter failure, second generator failure
• Root Cause: Pressure loss caused by failed high pressure seal in the pneumatic system

How can it be prevented from happening again?

• Proper maintenance
• Implement EOP to divert air (the ops team was new and did not know how to operate bypass)

Who’s fault is it?

• DC operator for not maintaining the pneumatic seals and for not ensuring adequate training
Transformer Core Saturation

- Dual cord critical load
- IT load served by dedicated 2N power system using DRUPS
- 3 Pole static transfer switches
- STS output feeds isolation transformer
- Isolation transformers feed IT load
- Data centre independently reviewed by several parties
- Evidence the entire commissioning system was flawed
- 100+ms utility event triggers DRUPS to start engines
- One of the DRUPS loses frequency control of output voltage during kinetic energy store operation
- DRUPS output frequencies drift apart
- Out of phase condition presented to STS inputs
- STSs transfer immediately to the alternate source
- Isolation transformer cores saturate
- Voltage output sags for 200 - 300 milliseconds
- IT load voltage drops to unsustainable level
- Critical load loses power

How could a 2N system experience both systems failing simultaneously?
Lessons Learned from Transformers Core Saturation

What failed?
• Power system failure leading to a complete data centre failure
• Financial services trading halted

Why did it fail?
• Voltage time integral exceeded isolation transformer saturation limit
• Causal factors: transient utility event, DRUPS lost frequency control
• Root Cause: Out of phase condition caused by incorrect STS settings

How can it be prevented from happening again?
• Enable STS delay transfer function
• Field test dynamic response to out of phase condition at STS input

Who’s fault is it?
• Design of this type of arrangement must always be cognizant of using STS delay transfer function
• Commissioning didn’t check STS delay setting or dynamically test STS out of phase input condition
Data centre has a high pressure gas fire suppression system

Requires activation of two separately zoned smoke detectors to trigger gas discharge

Evidence that gas discharge noise can damage or destroy computer hard drives

- High pressure gas discharge can exceed 135dBA
- HDD performance degradation commences at noise levels above 110dBA
- Failure of some types of HDDs as low as 118dBA
- Latest HDDs more susceptible to noise

Faulty piece of IT hardware causes smoke release into the data centre

Initially one then two smoke detectors go into alarm

Gas is discharged

HDDs fail
Lessons Learned from Fire Suppression Discharge
Sound Damages Hard Drives

What failed?

• Hard disk drives affecting multiple business processes

Why did it fail?

• Hard disk drives damaged by high noise level emitted by gas discharge nozzles
• Causal Factors: design of smoke detection system, noise level of high pressure discharge
• Root Cause: Inappropriate placement of discharge nozzles too close to HDDs

How can it be prevented from happening again?

• Use quieter nozzles
• Move nozzles away from HDDs such that noise <100dBA
• Use low pressure gas discharge
• Ensure smoke detectors are not too sensitive

Who’s fault is it?

• MEP engineers design of fire detection and suppression system is ultimately at fault
• High pressure fire suppression OEMs for not coming clean and notifying people of the problem
Residual Current Device Vulnerabilities to IT Loads

- Dual cord IT load
- IT load served by dedicated 2N power system
- Each final circuit protected by 30mA RCD
- Harmonic rich environment

- RCDs are imprecise e.g. allowable trip ranges of $0.5I_{\Delta N} - I_{\Delta N}$ or $0.11I_{\Delta N} - 2I_{\Delta N}$ depending on RCD type
- Research shows distorted current changes the RCD tripping threshold\(^1,2,3,4\)
- Aggregated earth leakage current - common cause of nuisance tripping – if server $I_{\Delta N} \approx 2mA$
- Scenario: 8 servers → 16mA → ok to trip a 30mA RCD

- Server power supply failure
- Server internal crowbar circuit generates high short to clear the internal fuse
- Transient in conjunction with the normal earth leakage causes RCDs to trip

\(^1\) Czapp S. The Effect of Earth Fault Current Harmonics on Tripping of Residual Current Devices. Analysis of the Residual Current Devices Independent Trip \(\text{Impact of Higher Order Harmonics on Tripping}^2\)

3. Yu Xiang, Wong X.H., Chen M.L. Tripping Characteristics of Residual Current Devices under
Lessons Learned from RCD Vulnerabilities to IT Loads

What failed?
- Power system failure leading to large scale data centre failure
- SLA breach

Why did it fail?
- RCD tripping characteristics not understood
- Causal factors: server power supply failure tipped local RCDs over trip threshold
- Root Cause: Use of incorrect type of RCDs

How can it be prevented from happening again?
- Don’t use RCDs in data centres unless local regulations demand it
- If forced to use them use with delay function and some harmonic immunity
- Continuously monitor aggregate leakage current

Who’s fault is it?
- Design error - must always be cognizant of electromagnetic RCD characteristics
- Operator should ensure aggregated leakage is monitored
Uncoordinated Application Chains

- Event takes places at investment back during mid afternoon trading
- Active-Active Data Centres - live applications running in both DCs
- Electrical fault in one CRAC unit
- Two under floor smoke detectors simultaneously go into alarm
- Fire suppression discharges gas
- EPO power shut down to all IT equipment and CRAC units
- Major mainframe production and storage impact
- Network connectivity between data centres lost
- All IT components replicated for fail over
- Some primary and secondary application and databases are located in the same data centre
- +2 hours CRAC units returned to service
- +3 hours power restored to all PDUs
- +5 hours IT restoration and communications reestablished with mirror site
Lesson learned from Uncoordinated Application Chains

What failed?
• Power and cooling leading to failure of one side of a mirrored data centre
• Financial services trading halted and cannot settle trades

Why did it fail?
• Hot and Standby applications located incorrectly one data centre
• Causal factors: CRAC unit electrical failure, gas discharge, auto power down of PDUS and CRACs
• Root Cause: Active and fail over applications located in the same data centre

How can it be prevented from happening again?
• Map business process and application chains down to physical technology component level
• Implement process for ongoing change management linking application and IT infra management
• Develop integrated IT delivery process consistent with business process requirements
• Limit proliferation of divergent application chains resiliency models

Who’s fault is it?
• Application services management
• MEP engineers design fire detection and suppression system
Common Recurring Failures

- Generators fail to start
  - Batteries - ops
  - Air lock in Fuel - design
  - Contaminated Fuel - ops

- Uncoordinated Circuit Protection – design / commissioning
- Loose connections - Switchgear / Busbar – contractor
- UPS battery failure - OEM
- PLC logic – design
- Water leak
- Standard Operations Switching errors
- Maintenance Operations Errors
- Design Errors e.g. controls systems
The effect of equipment failures, human error and Acts of God can be visualised using the “Swiss Cheese” model.

Defensive control layers try to minimize occurrence of unplanned downtime.

Unplanned outages occur with the alignment of holes in successive control layers.

Holes are a combination of:
1. Equipment Failure
2. Human Error
3. Acts of God

Source: “Swiss cheese” model (Adapted from Reason, 1997)
At some point all equipment will fail.

Equipment failures can be modelled mathematically.

Reliability modelling uses known component failure rates ($\lambda$) and Monte Carlo simulations to predict the overall system reliability and availability.

Used by mission critical industries e.g. aviation, nuclear etc.
Example Pump Power Failure Modes

<table>
<thead>
<tr>
<th>Device Type</th>
<th>Failure Mode</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Battery, lead-acid</td>
<td>Degraded output</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td>Short</td>
<td>.20</td>
</tr>
<tr>
<td></td>
<td>Intermittent output</td>
<td>.10</td>
</tr>
<tr>
<td>Cable</td>
<td>Short</td>
<td>.45</td>
</tr>
<tr>
<td></td>
<td>Earth leakage</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>Open</td>
<td>.19</td>
</tr>
<tr>
<td>Circuit breaker</td>
<td>Opens without stimuli</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td>Does not open</td>
<td>.49</td>
</tr>
<tr>
<td>Connector/connection</td>
<td>Open</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>Poor contact/intermittent</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>Short</td>
<td>.16</td>
</tr>
<tr>
<td>Electric motor, AC</td>
<td>Winding failure</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>Bearing failure</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>Fails to run after start</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>Does not start</td>
<td>.18</td>
</tr>
<tr>
<td>Pump, centrifugal</td>
<td>No output</td>
<td>.67</td>
</tr>
<tr>
<td></td>
<td>Degraded output</td>
<td>.33</td>
</tr>
</tbody>
</table>
Because we are only interested in failed states, the number of combinations is given by \( C = F^2 - 1 \). So for a device that has 2 failure modes, the number of combination is \( 2^2 - 1 = 4 - 1 = 3 \).

In the example, the number of combinations of failed states is given by the product of the QC items

\[
F = (Q_1 \times Q_2 \times Q_3 \times Q_4)
\]

\[
F = 16 \times 6 \times 8 \times 15
\]

\[
F = 11,520 \text{ theoretical combinations of failed states}
\]
Many of the theoretical combinations of failure modes are unrealistic because they are:

1. **Highly Unlikely** to all co-exist
   e.g. a cable simultaneously having o/c and s/c

   or are

2. **Mutually Exclusive**
   e.g. motor can not simultaneously fail to start and fail to run after start

This significantly reduces the number of possible combinations of failure modes
Example Pump Power Failure Modes

The number of practically possible failure combinations is still given by:

\[ F = (Q_{C1} \times Q_{C2} \times Q_{C3} \times Q_{C4}) \]
\[ F = 14 \times 4 \times 3 \times 8 \]

\[ F = 1344 \] practically possible combinations of failed states

<table>
<thead>
<tr>
<th>DEVICE</th>
<th>FAILURE MODES</th>
<th>TESTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>CABLES</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>CIRCUIT BREAKERS</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>STARTER</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>MOTOR</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

When the circuit is initially powered up with the pump on load, after a run-in period, the majority of device combination failure modes are proven.

The remaining combinations can then verified through standard electrical safety tests.
• We learn by experiencing failure ourselves or the shared experience of others
• As we learn we descend the curve
• Every organization should establish where it is on the ULC

Universal Learning Curve

![Diagram showing the Universal Learning Curve (ULC)](image)

Source: Duffey & Saull – Managing Risk The Human Element
Human Bathtub Curve

- Double exponential
- \( p(\xi) = 1 - e^{-\left\{\lambda - \lambda_m\right\}/k - \lambda(\xi_0 - \xi)\} \)

where
- \( p(\xi) \) is the error probability
- \( \lambda_o \) is minimum attainable failure rate
- \( \lambda \) is the failure rate
- \( \xi_0 \) is initial experience level
- \( \xi \) is current experience level

The initial or first event has a purely random (Bayesian) occurrence

Eventually, when very large experience is attained, we climb up the curve again because we must have an event

We descend the curve by learning from experience thus reducing the chance or risk

The bathtub bottom or minimum risk is eventually achieved

Increasing Experience (for the homo-technical system)
<table>
<thead>
<tr>
<th><strong>FAILURES</strong></th>
<th><strong>WHO?</strong></th>
<th><strong>WHAT?</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>System Design</td>
<td>Consulting Engineer</td>
<td>Human</td>
</tr>
<tr>
<td>Systems Validation</td>
<td>Commissioning Engineer</td>
<td>Human</td>
</tr>
<tr>
<td>Equipment Design</td>
<td>OEM</td>
<td>Human</td>
</tr>
<tr>
<td>Equipment Failure</td>
<td>No one</td>
<td>Machine</td>
</tr>
<tr>
<td>Operator Error</td>
<td>Ops</td>
<td>Human</td>
</tr>
<tr>
<td>Installation Error</td>
<td>Contractor</td>
<td>Human</td>
</tr>
<tr>
<td>Maintenance Oversight</td>
<td>Ops</td>
<td>Human</td>
</tr>
<tr>
<td>Natural Disaster</td>
<td>Act of God</td>
<td>Various</td>
</tr>
</tbody>
</table>
Concluding Lessons Learned

- Many data centre failures caused by consulting engineer and commissioning
  - Lack of mission critical knowledge
  - Time pressure
  - Poor operations and maintenance documentation
  - Lack of handover training (design – operations)

- How do we minimise operator human error rate?
  - Continuous training
  - Instrumentation and ergonomic HMI interface
  - Learning through training is most important

- Senior Management often complicit in failure due to focus on operating costs vs. reliability → failure

- We learn from understanding causes of failure and how to avoid them

- When the data centre industry shares failure information overall reliability will improve
Thank You

Ed Ansett, i3 Solutions Group
9th November 2016