5G Core Network Reliability

Thoth Advisory Perspective on Network Resilience

Contents

Executive Summary

5G adoption has experienced the fastest growth of any previous cellular generation with 260 commercial 5G networks deployed as of June 2023 and over 750 5G smartphone models available to users. In terms of the number of global 5G connections, 5G broke through the 1 billion mark in December 2022 and at the end of June 2023 surpassed 1.2 billion.

The power of 5G – ultra-low latency (<10 ms), 1 million IoT devices per square km, 10 Gbps downlink (DL) in 3GPP Rel-18, network slicing and so on - is only possible with a complex 5G Core architecture that manages and directs all aspects of connectivity and subscriber management. The 5G Core interfaces to the Business Support Systems (BSS) for charging and billing and the 5G Core routes data to the core routers for connectivity to the Internet. Many MNOs are operating both 4G and 5G networks with voice calling often times falling back to the 4G network.

Network reliability signifies the ability of a network to minimize the scope and frequency of network incidents, continue operations while under pressure and recover as quickly as possible. *Network resilience* can be defined as the ability of the network to provide and maintain an acceptable level of services in the face of various faults and challenges to normal operation. The 3GPP 5G specifications include functionality to improve reliability and availability but the sheer complexity of the 5G Packet Core with 10+ Virtual Network Functions means that reliability is much more complex with 5G and will invariably require technology innovation up and above the 3GPP basic specifications.

In this Whitepaper, we discuss some of the innovations that are being developed by the industry to address 5G Core reliability. We begin by looking at the situation with global network incidents that have been reported in the public domain around the globe. We then provide a brief review of the theory of reliability and its mathematical probability foundations. We then take a closer look at just what are "network fault", how to detect them and how to measure parameters that indicate faults. Empirical evidence has shown that there can be many root causes for signal storms - a link going offline from cutting a cable in the data center, to software upgrades gone awry, to physical or logical links within the 4G and 5G core failing and causing a knock-on effect. As a result, a strategy for network reliability detection, measurement and remedies is of paramount importance. A successful strategy, which includes topology, ensures that no single network Element (NE) will become a single point of failure.

The industry has learned a lot in the past few years ever since 4G LTE was commercialized and then when 5G entered the market in 2018/2019. Based on industry insights, we propose in this WP a strategy for detecting and managing network faults in the 5G network. We thus propose a 3-6 layer architecture and six network actions that can be used to tackle faults and resolve the root cause and restore operations. Finally, we discuss a framework for defining the health of the network.

The Challenge: Reducing Network Outages

How often do network outages occur and what is their impact?

The consequences of disruptions of the internet are increasingly severe, and threaten the lives of individuals, the financial health of businesses, and the economic stability and security of nations and the world. With the increasing importance of the Internet, so follows its attractiveness as a target for attackers, whether they be recreational or professional hackers, terrorists or those intent on information warfare.

It is generally accepted that the current Internet is not as resilient, survivable, dependable, and secure as industry and commerce require. Resilience is a fundamental requirement of any future network design and must be built into the components as well as in the network upper-layers. Communications networks are constructed as a multilevel stack of infrastructure, protocols, links and nodes, topology, routing paths, end-to-end transport and APIs. The ***resilience*** of each of these levels in the stack provides a foundation for the next level. Multi-level resilience is implemented by three critical strategies:

1. Redundancy for fault tolerance
2. Diversity for survivability
3. Connectivity for disruption tolerance

Figure 1 illustrates some recent examples of serious network telecom service provider (telco) outages that have occurred around the globe. A partial list of some the root causes of the global disruptions include:

- Interconnection systems between different mobile operators
- Rogue RAN site
- 4G Evolved Packet Core software upgrade
- 3G packet switch
- FTTH routers
- LTE EPC software upgrade
- Voice switching upgrade creating a signal burst
- Cable cut in data center

The consequences of serious outages, of which the more extreme cases lasted 8 - 48+ hours, has created financial losses, hurt the telcos' brand, and even incurred financial penalties being imposed by regulators.

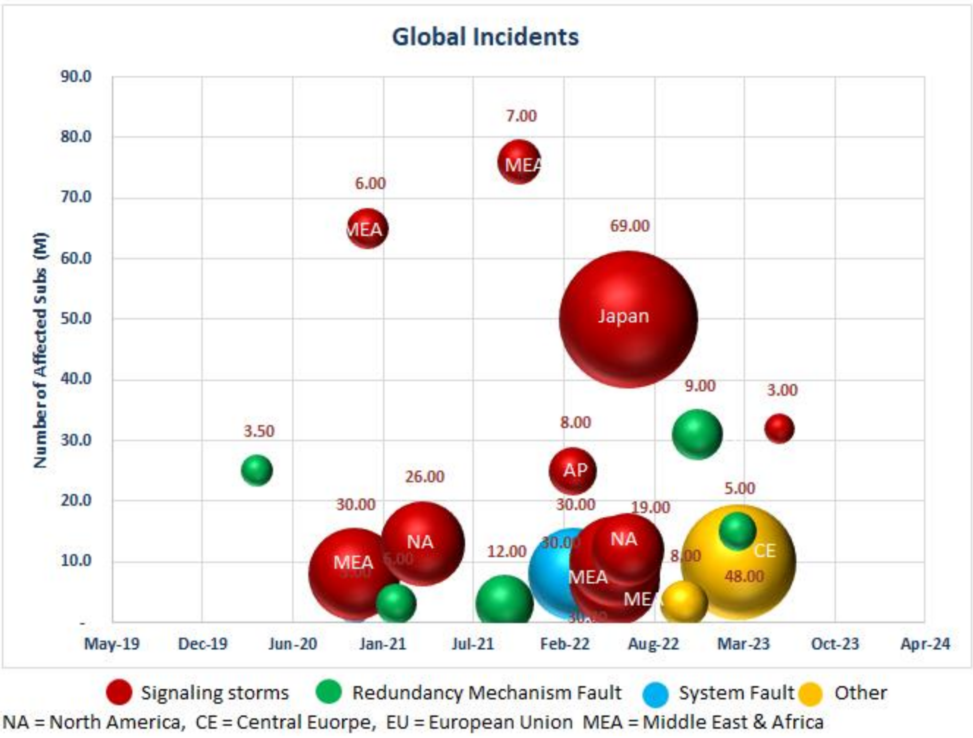
Why? Because the internet is now the lifeblood of commerce and impacts everyone's daily life activities. Oftentimes, the outage starts in one city or state but as the downtime hours progress, the disruption can spread to multiple cities and in some cases nationwide. Some operators have been forced to compensate customers for mass outages due to the inconvenience caused.

The U.S. has stipulated that communications equipment outages reach 900,000 subscriber minutes must be explained to Congress and fined.

Some noteworthy publicly reported incidents and data on incidents would include the following:

- In 2011, CA conducted a survey on 200 North American enterprises (with a total sales revenue of US$26.5 billion) and found that the average annual loss caused by IT faults caused by the company's annual loss was US$30 million. Each company lost an average of US$150,000 per year, and the loss was US$5.6 million per minute. (The average annual interruption duration of each company was less than 30 minutes.
- On August 17, 2013, Google broke down for 5 minutes, causing a 40% drop in global network traffic and a reported significant financial losses from that incident.
- In February 2010, a serious fault occurred on 3G equipment of Telecom New Zealand, affecting 200,000 subscribers. Telecom New Zealand payed out more than $10 million to its customers.
- On April 21, 2009, a vendor providing HLR was faulty on the T-Mobile network. As a result, 40 million subscribers could not make calls or send short messages for four hours.

Figure 1:Recent global network incidents



Note: Bubble size is the number of hours of downtime.

Source: Public Sources

Business Continuity Management

The 5G network service continuity encompasses both digital and physical systems such as the 5G Core, optical transport, aggregation access network, vRAN, edge cloud, cloud connectivity, and data center routing and switching. 5G introduces state of the art concepts such as virtual RAN, cloud-native 5G Core architecture, network slicing and ultra-low latencies. As a consequence the added complexity that comes with these new functionalities necessitates stronger resilience and fault tolerant design in both the hardware, application software and communications software layers. The number of 5G Call Detail Records (CDRs) generates is >6X that in LTE so the real-time OSS and the BSS systems need to have resilient storage systems.

Rapid disaster recovery (DR) and back up strategy are critical to maintain business continuity. Recovery Point Objectives (RPO) and Recovery Time Objectives (RTOs) are two vital metrics that the network must adhere to. RPO defines the acceptable amount of data loss for the telco and RTO is the recovery window (i.e. the service loss) for systems and applications after an outage. Data centers can be configured into active-active and active-passive across local and remote Disaster Recovery locations. In active-active topology (RPO=0, RTO=0), two DCs provide concurrent service by ensuring the technology stacks work together cohesively. In active-active design, load balance is maintained by using distributed controller nodes and software-defined architecture for traffic load sharing. Active-active requires redundancy design in the optical transport as well including multiplexers, GPON OLTs, DCI network devices and 1+1 protection schemes.

Telcos today are addressing the reliability of storage systems by selecting purpose-build backup appliances (PBBAs) which provide various features including high-data deduplication 10-65:1. Rapid remote office backup, and protocol translation when transferring data to cloud repositories. Telcos are also moving to migrate more and more of their storage systems for their internal operations from HDD to All Flash Arrays (AFA) with the added benefit of lower latency and better MTBF.

The new GenAI applications will need specialized and separate data center infrastructure with rack power density 3-4X those found in today's telco and hyperscaler DCs. Thus, as work loads start to include a GenAI component the DR, RPO and RTO parameters will need to be fine tuned for GenAI as well as for traditional workloads.

Figure 2 summarizes business continuity aspects. The interface between human and system affects the stability of system operation. Therefore, we need to consider reducing human participation (through automation), standardizing operations, and taking remedial measures for mis-operations.

- **Environment:** Changes in the operating environment of the device affect the normal operation of the system. Therefore, the environment needs to be monitored at any time and the system needs to respond to major and catastrophic events.
- **Network:** The network is an information channel in distributed deployment. Data consistency, availability, and partition isolation (CAP) must be considered.
- **Software:** detects hardware faults and automatically processes them to ensure data consistency. Note that common-mode faults in software cause avalanche effects under certain conditions. Fault isolation can be used to control the avalanche effect.
- **Hardware:** Various hardware faults can occur anytime and anywhere in the system. Operators need to understand these faults, detect them, and automatically handle them. Common hardware monitoring includes:
  - Fan modules in redundancy mode: dual power supplies in redundancy mode
  - Onboard monitoring software: BMC, dual mirroring, high reliability
  - Memory reliability technologies such as ECC, hot-swappable hard disks, and various RAID modes This feature improves disk reliability. Supports online scheduled fault detection and warning of disks, NICs, CPUs, and memory. Supports subhealth indicator detection and smart scoring. Provides detailed system monitoring for servers.
  - Monitor system information such as CPU, memory, hard disk usage, and temperature in real time. Power-on self-check: 2285RAID1068/1078/1064; battery health, fan speed adjustment, board isolation and maintenance mode. Further, industrial-grade component selection to prevent mis-insertion, CPU single-core fault detection failure analysis and DPA derating design, thermal simulation, noise reduction, EMC, and vibration. Environment compatibility design and test related certification HALT/HASS/safety/aging/mass production data reliability spot check.

Figure 1: Business/Service Continuity Management

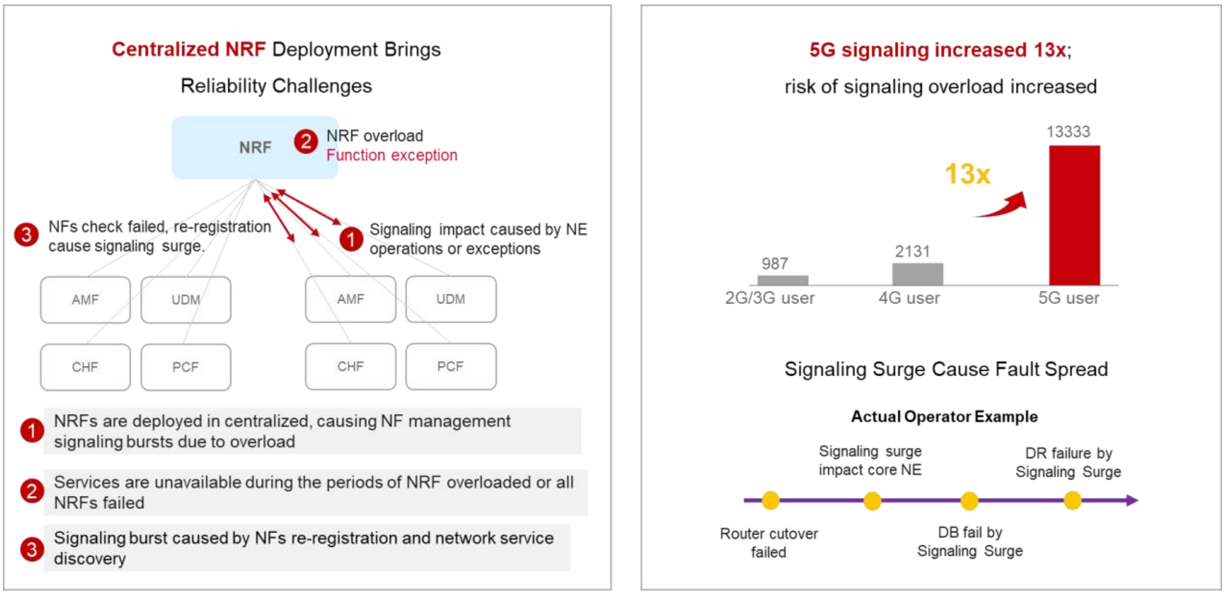
Factors	Fault risk	Countermeasures
Hardware	Hard disk, NIC, memory, CPU, power supply, or fan fault	Hardware status/performance/resource detection/disk scoring/fault isolation
Software	Process fault/Resource exhaustion/Split-brain/Data loss/Data inconsistency/Service congestion/	Redundancy design, process monitoring, resource monitoring, alarm design, correlation analysis, hierarchical reset, and flow control
Network	Network storm, network QoS deterioration, and network interruption	Heartbeat detection, multi-plane isolation, redundancy design (GPON, DCI, multiplexers), multipathing, fast failover, subhealth monitoring, and policy-based fault locating
Environment	Power outage/air conditioner overheating/rat infestation/natural disasters	Environment parameter monitoring, intelligent fan, emergency plan, routine drill, and disaster recovery
People	Mis-operation, configuration error, and malicious attack	Automatic management, validity check, configuration rollback, silent operation, and security measures

Source: Thoth Advisory

The 5G Core Architecture

The complexity of the 5G Core and protocols creates significant network reliability challenges. First, the volume of 5G signalling as compared to 4G LTE is 13X and consequently the risk of signalling overload increases substantially. An example of how this can happen consider a signalling impact caused by a Network Element (NE), such as a router, experiencing an exception. The NRF (Network Repository Function) would then overload creating its own exception and the VNFs check would fail, and repetitive re-registration then leads to a signal surge.

Figure 2: 5G Reliability Challenges increase with 5G NR


Centralized NRF Deployment Brings
Reliability Challenges
NRF
2 NRF overload
Function exception
3 NFs check failed, re-registration cause signaling surge.
1 Signaling impact caused by NE operations or exceptions
AMF
UDM
CHF
PCF
AMF
UDM
CHF
PCF
1 NRFs are deployed in centralized, causing NF management signaling bursts due to overload
2 Services are unavailable during the periods of NRF overloaded or all NRFs failed
3 Signaling burst caused by NFs re-registration and network service discovery
5G signaling increased 13x;
risk of signaling overload increased
13333
13x
987
2131
2G/3G user
4G user
5G user
Signaling Surge Cause Fault Spread
Actual Operator Example
Signaling surge impact core NE
DR failure by Signaling Surge
Router cutover failed
DB fail by Signaling Surge

Note: AMF=Access and Mobility Function PCF = Policy Control Function, UDM = Unified Data Management, CHF = Charging Function, NRF = Network Repository Function

Source: Thoth Advisory, HPE

5G Private Wireless Networks (PWN) need high reliability

5G PWNs are a special case of 5G networks where the network is built for a specific campus, utilities distribution network, stadium, manufacturing facility, industrial complex, shipping port or mining camp. 5G PWN are being deployed in:

- Factories – discrete and process manufacturing
- Warehouses (these might be attached to factories as well) and logistics centers
- Utilities (electricity, gas, and oil pipelines) substations and distribution lines
- Mining camps
- Oil and Gas production (onshore, terrestrial and offshore)
- Railway lines and metro stations (these can be underground as in the case of subways)
- Campuses – business complexes and industrial parks, schools
- Sports venues
- Government building complexes
- Merchant shipping which includes cargo ships, oil tankers and cruise ships

In a typical 5G PWN in a factory, for example, a 5G gNodeB centralized unit (CU) and a coordinated multipoint (CoMP) server drive multiple gNodeB transmission and reception points (TRPs), the latter of which are placed around the factory to provide full area coverage. In general, the number of TRPs is equivalent to the number of Wi-Fi 6 access points (APs) in order to achieve the same coverage. 5G creates spatial diversity with redundant communication paths across both indoor and outdoor environments using the CoMP server. The private 5G network can be built without CoMP, but CoMP does present a convenient way of improving the network efficiency in the indoor environment.

Non Standalone and Standalone (SA) PWNs are being built but the preferred direction is to build with SA because of the additional functionality and lower-latencies and network slicing. For business operations PWNs can become vital to the operations and reliability, availability and resilience take on the utmost important because downtime will directly impact the bottom line.

There is an additional aspect to PWNs which might be easily replicated in the public network areas and that is the density of devices and the traffic densities. It only takes a dozen Automated Mobile Robots (AMR) in a factor floor with 4K streaming to wreck the 5G network performance and stakeholders are learning that total bandwidth capacity must be as high as possible in order to ensure high availability of the manufacturing lines. In some cases the 5G core is implemented on a cloud platform but many manufacturing companies will prefer to have all 5G Core components that affect real-time processes to be installed locally with the cloud serving a more back-up role as well as a platform to run data analytics of the date retrieved from sensors.

All of the techniques for measuring network congestion and signal surge prevention apply equally importantly to PWNs and some cases PWNs might even be more challenging because of the device density: a single mining site might have 1000s of sensors, IP sensors and tools all connected to a local LAN network.

Basic Concepts of Network Resilience

Resilience

Network reliability signifies the ability of a network to minimize the scope and frequency of network incidents, continue operations while under pressure and recover as quickly as possible.

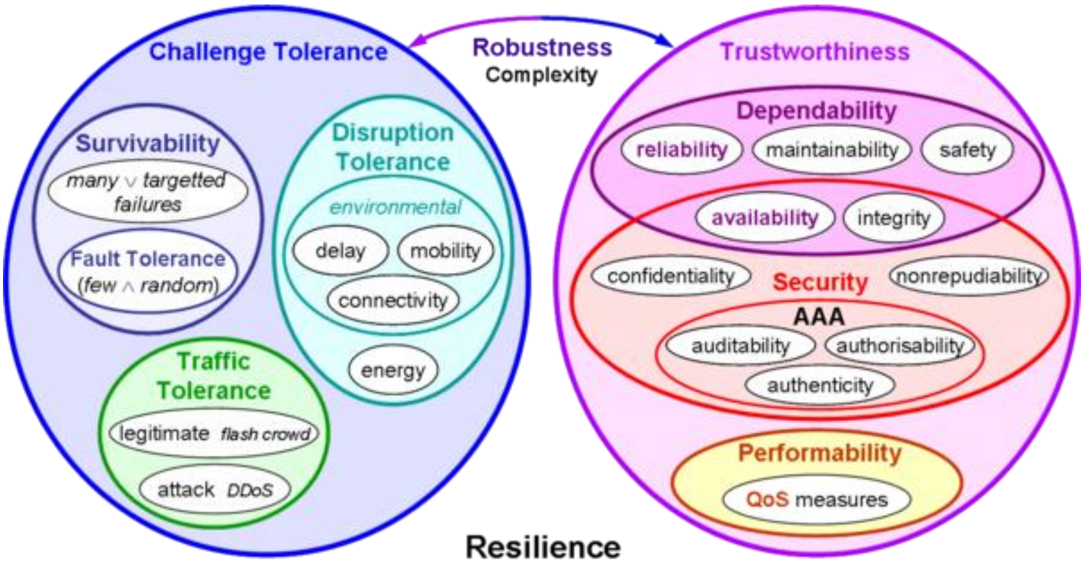
Network Resilience can be defined as the ability of the network to provide and maintain an acceptable level of services in the face of various faults and challenges to normal operation (Sterbenz, J.P. et al. *Computer Networks* 2010).

Network Topology. The ability of a network to survive/withstand component failure or external hostile actions is linked invariably to the network topology as well as the individual component/subsystem failure rates (Mean Time Before Failure - MTBF). A networks' topology is made of connected nodes that link network elements (NEs) that can originate or receive messages. Most links are bidirectional but some may also be unidirectional (such as IoT up-streamed data). Node-positioning and cost considerations are important aspects of topological analysis with respect to network reliability. *Node connectivity* refers to a link path existing between two nodes such as between optical transport nodes. Graphical properties that go into the node analysis include degree distribution, shortest path distribution, and link length distribution.

Resilient networks remain in normal operation in the face of challenges.

Network survivability is the ability of the network to withstand component failure, cyberattacks and subversion and is a function of the survivability of network topology and individual nodes and NEs. Ellison et al. in 1997 provided a definition of survivability as "*the capability of a system to fulfil its mission, in a timely manner, in the presence of attacks, failures or accidents.*" Threats includes targeted attacks or large-scale natural disasters resulting in many failures, in addition to the few random failures covered by fault tolerance. Survivability is thus a superset of fault tolerance but a subset of resilience. Pioneering work in survivability analysis is based on the Three-tuple (triple) which is a function of Reliability, Resilience, and Recognition. Survivability research dates back to the late 1990s which sought to address critical infrastructures in telecommunications, power grid, transportation and bank payment systems.

Figure 3: Network Resilience



Source: ResiliNets Project, 2015

Fault Tolerance is the ability of a system to tolerate faults such that service failures do not result. Fault tolerance generally covers random single or at most a few faults, and is thus a subset of survivability, as well as of resilience.

Disruption Tolerance is the ability of a system to tolerate disruptions in connectivity among its components. Disruption tolerance is a superset of tolerance of the environmental challenges: weak and episodic channel connectivity, mobility, delay tolerance, as well as tolerance of power and energy constraints.

Traffic Tolerance is the ability of a system to tolerate unpredictable offered load without a significant drop in carried load (including congestion collapse), as well as to isolate the effects from cross traffic, other flows, and other nodes. The traffic can either be unexpected but legitimate such as from a flash crowd, or malicious such as a DDoS attack.

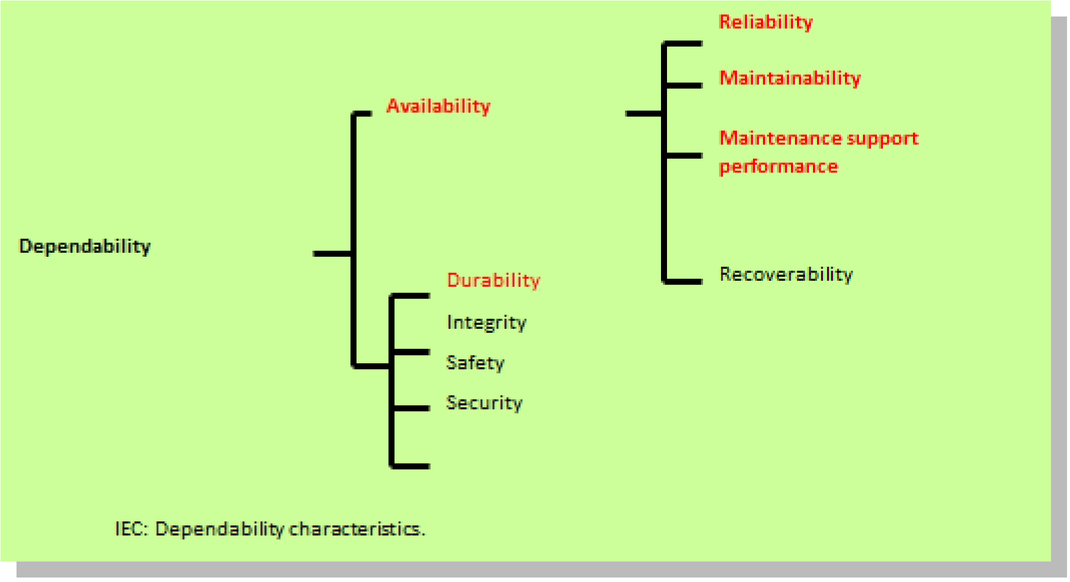
Trust

Security is the property of a system and measures taken such that it protects itself from unauthorised access or change, subject to policy. *Security properties* include AAA (auditability, authorisability, authenticity), confidentiality, and nonrepudiation. Security shares with dependability the properties of availability and integrity.

Performability is the property of a system such that it delivers performance required by the service specification, as described by QoS (quality of service) measures.

Dependability is the property of a system such that reliance can justifiably be placed on the service it delivers. It generally includes the measures of availability (ability to use a system or service) and reliability (continuous operation of a system or service), as well as integrity, maintainability, and safety. Dependability characteristics include availability and its influencing factors (reliability, recoverability, maintainability, maintenance support performance) and, in some cases, durability, integrity, safety and security. Dependability is often used descriptively as an umbrella term for the time-related quality characteristics of a product or service. Specifications for dependability characteristics typically include: the function the product is required to perform; the time for which that performance is to be sustained; and the conditions of storage, use and maintenance. Requirements for safety, efficiency and economy throughout the life cycle may also be included.

Figure 4: Definition of Dependability



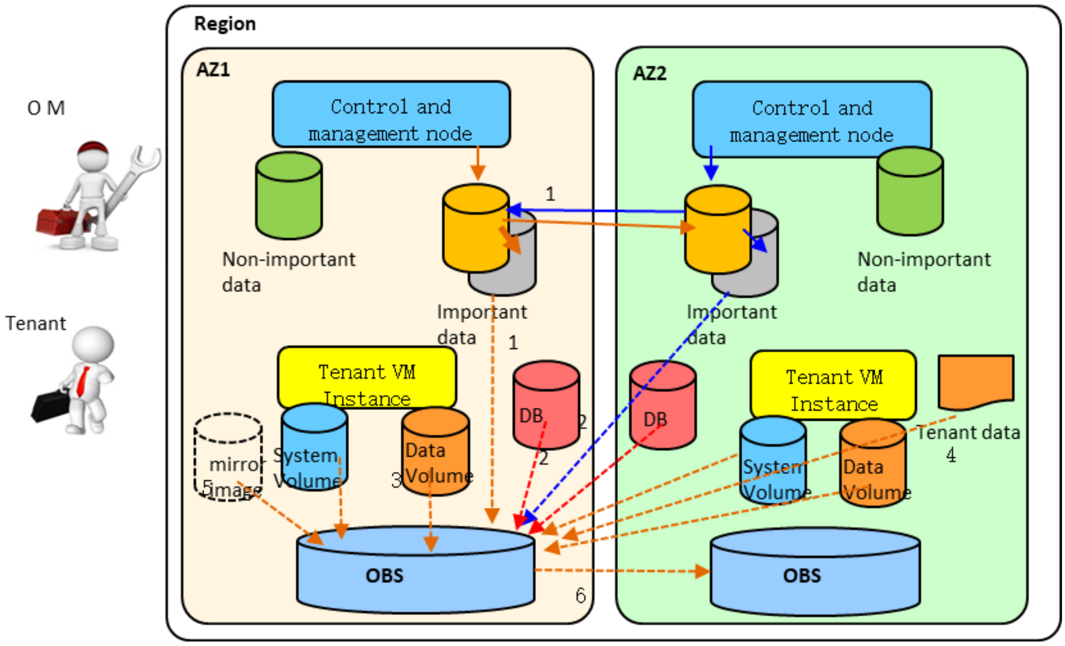
General understanding of the international standards community

Source: Thoth Advisory

Durability. Durability reflects the probability of data loss and is the first quality attribute of storage products. Durability takes precedence over availability, performance, cost, and scalability. Figure 6 shows tenant-related data durability:

- Data in the OS cache of a VM cannot be ensured. When the OS breaks down, data is lost.
- EVS block storage uses three copies, and the durability can reach five to seven nine nines. (depending on the disk type, SATA/SAS/SSD), while the durability of common commercial disks is only one nine nines.
- OBS uses EC codes (+3 redundancy) in the AZ, and the data durability reaches 99.999%.
- OBS supports cross-AZ replication, providing data durability up to 11 nines.

Figure 5: Tenant-related data durability



Source: Thoth Advisory

Figure 6: Multi-Level Data Durability principles in the public cloud

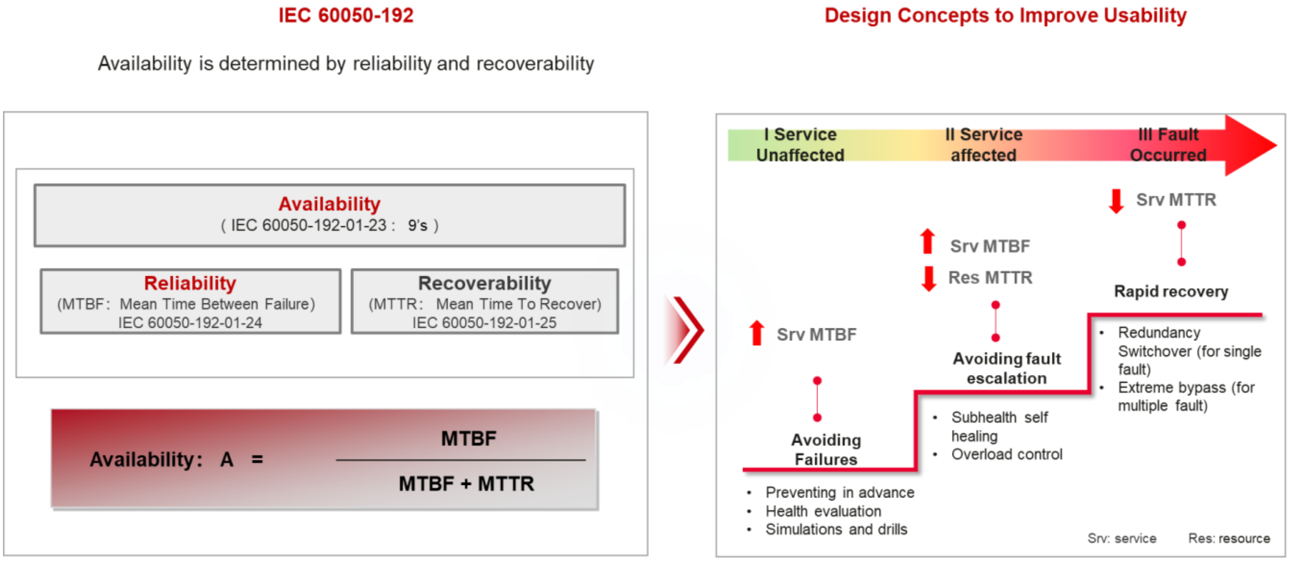
Manage Data **- O&M personnel provide different backup mechanisms based on data importance.**	Important data: System Configuration Data Tenant authentication information Tenant VM information Tenant disk information	Important data must be highly durable. In addition to local data active/standby redundancy, the following methods can be used: 1. Management data is backed up to OBS or FTP server (RPO < 1 hour). 2) Cross-AZ HA, meeting the infrastructure service recovery requirements in disaster scenarios.
	Non-important data: Monitoring information Alarm statistics	If the data volume is large but the durability is not required, data can be periodically backed up to OBS or FTP server (RPO < 1 day).
Tenant data **- Tenants use different backup services based on data importance and cost.**	Disk Data	Disk data uses three copies, and the data durability can reach five to seven nines, which is much higher than the durability of common commercial disks (2 nines). If tenants require higher data durability, they can use the following methods: 1. VBS is used to back up key disks to OBS. 2) Cross-region backup replication, implementing cross-region data recovery
	Important documents	1. OBS The object storage service (OBS) stores key data in OBS. (2) OBS cross-AZ data replication, further improving data durability
	RDS database	1. Manually or automatically backed up data to OBS by RDS. (2) Cross-AZ RDS data replication
	Mirror Data	Image data is stored in OBS, and the durability reaches 7 to 11 nine nines.

Source: Thoth Advisory

Network Resilience Design Principles

An essential aspect of resilient network design is to understand how the networks behave under various challenges. To analyse network resiliency, network engineers model the challenges that disrupt the normal operation of the network, and this can be done through the use of simulation scenarios against ***n*** networks for ***c*** challenges.

Figure 7: Constructing Network High Availability Based on Classical Theory



IEC 60050-192

Availability is determined by reliability and recoverability

Availability
(IEC 60050-192-01-23 : 9's)

Reliability
(MTBF: Mean Time Between Failure)
IEC 60050-192-01-24

Recoverability
(MTTR: Mean Time To Recover)
IEC 60050-192-01-25

$$\text{Availability: } A = \frac{\text{MTBF}}{\text{MTBF + MTTR}}$$

Design Concepts to Improve Usability

I Service Unaffected — II Service affected — III Fault Occurred

Srv MTBF
Avoiding Failures
- Preventing in advance
- Health evaluation
- Simulations and drills

Srv MTBF
Res MTTR
Avoiding fault escalation
- Subhealth self healing
- Overload control

Srv MTTR
Rapid recovery
- Redundancy Switchover (for single fault)
- Extreme bypass (for multiple fault)

Srv: service Res: resource

Source: Thoth Advisory

3-Layer Resilience Theory – ResiliNets

The *ResiliNets* initiative began in 2013 led by Lancaster University (UK) and the University of Kansas. ResiliNets is actually an umbrella for a number of projects in resilient future Internet architecture including

- ***Post-Modern Internet Architecture (PoMo)*** which studies heterogeneous networks
- ***Great Plains Environment for Network Innovation (GpENI)*** which investigates programmable infrastructure
- ***ANA (Autonomic Networking Architecture).*** It aims to understand and improve the resilience and survivability of computer networks, including the Internet. The ResiliNets project aims to understand and improve the resilience and survivability of computer networks, including the global Internet, PSTN (public switched telephone network), SCADA (supervisory control and data acquisition) networks, mobile ad hoc networks and sensor/IoT networks. The ResilieNets project also developed two communications protocols: *ResTP: Resilient Composable Multipath Transport Protocol* and *GeoDivRP: Geodiverse Mtulipath Routing Protocol*.
- ***ResumeNet*** *(See Below)*

ResiliNets Actions: D^{2}R^2 Detect, Defend, Remediate, Recover

The ResumeNet project (*Resilience and Survivability for Future Networking: Framework, Mechanisms, and Experimental Evaluation*), under the ResiliNets umbrella, hosts original research work that aims to systematically embed resilience into the future Internet. Participants include the University of Kansas, Lancaster university (UK), Techniche Universiteit Delft, NEC laboratories, and France Telecom Orange Labs. The project developed metrics, classes of network resilience and policies and ways to negotiate them.

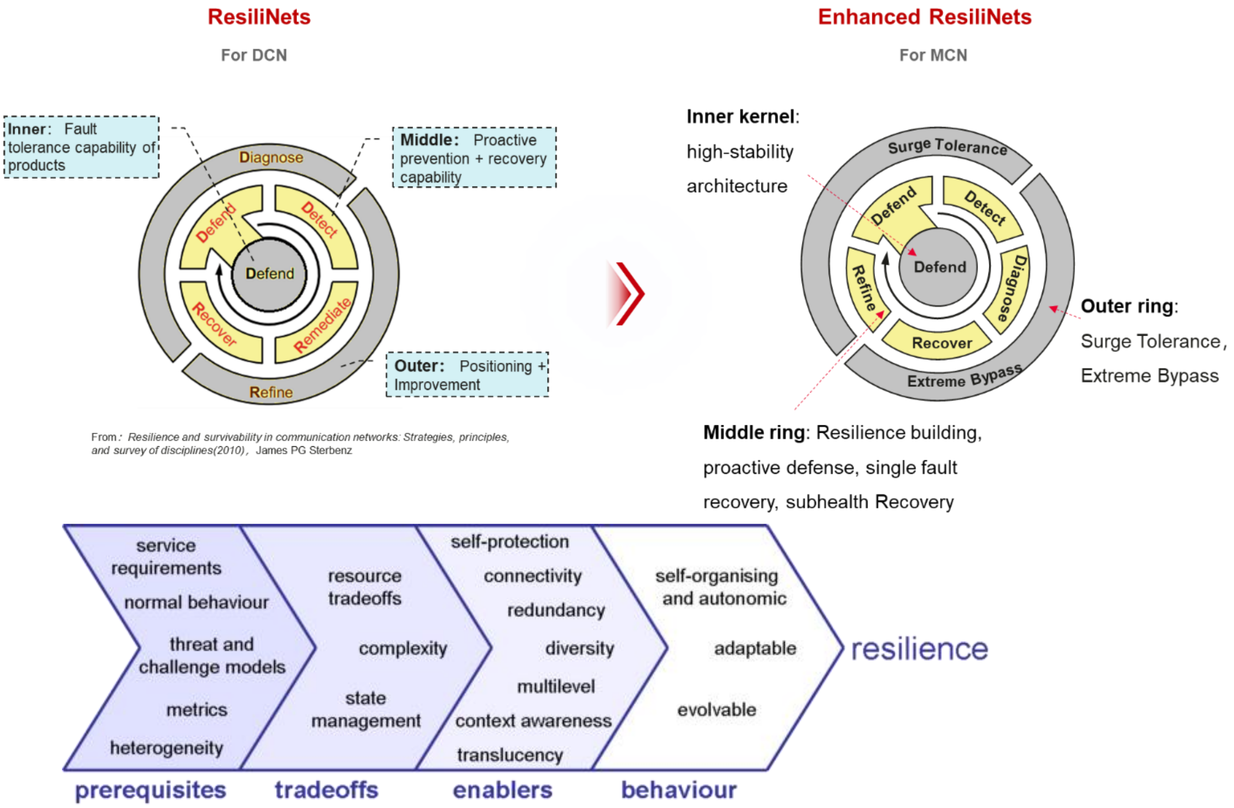
The Resilience Strategy in ResiliNets can be written as $\mathbf{D^2R^2}$ + **DR** (See Figure 3) where $\mathbf{D^2R^2}$ is a real-time loop consisting of four phases of operation in every network subsystem and protocol: **active defence** resists attacks and challenges on the network using mechanisms such as *filtering* on known threat signatures. When challenges do manage to penetrate the network, context-aware detection mechanisms trigger adaptive **remediation** such as *dynamic rerouting* and *walling off* compromised subsystems. Finally, **recovery** mechanisms and infrastructure redeployment are used to bring the network back to normal operation. The outer **DR** background loop is used to diagnose the root cause behind the penetration of the network and to perform analysis of the entire inner loop operation in order to refine future behaviour for improved $\mathbf{D^2R^2}$ operation *(Source: J. Sterbenz et al.2013)*

- The first **D** = Defend against challenges and threats to normal operation and includes both passive and active defense. A passive defense resists challenges to the network such as redundant, diverse topologies. The idea here is make the network as resistant as possible to challenges.
- The second **D**=Detect since inevitably a network will be threatened and it must be able to detect this automatically when an adverse event or condition has occurred
- The First **R**= remediate any damage to minimize the overall impact, and finally will recover as it repairs itself and transitions back to normal operation.
- The second **R**=Recover to original and normal operation
- **DR** = The Background loop which stands for **D**iagnose the fault that was the root cause and **R**efine future behavior. DR consists of diagnosing any design flaws that permitted the defences to be penetrated, followed by a refinement of network behaviour to increase its future resilience.

With the above two-phase strategy, $\mathbf{D^2R^2}$ + **DR,** it is then possible to derive a set of design principles leading to resilient networks:

- **Prerequisites:** {service requirements, normal behavior, threat and challenge models, metrics, heterogeneity in mechanism, trust, and policy}
- **Tradeoffs:** {resource tradeoffs, complexity, state management}
- **Enablers:** {security and self-protection, connectivity, redundancy, diversity, multilevel, context awareness, translucency}
- **Behavior:** {self-organizing and autonomic, adaptable, evolvable}

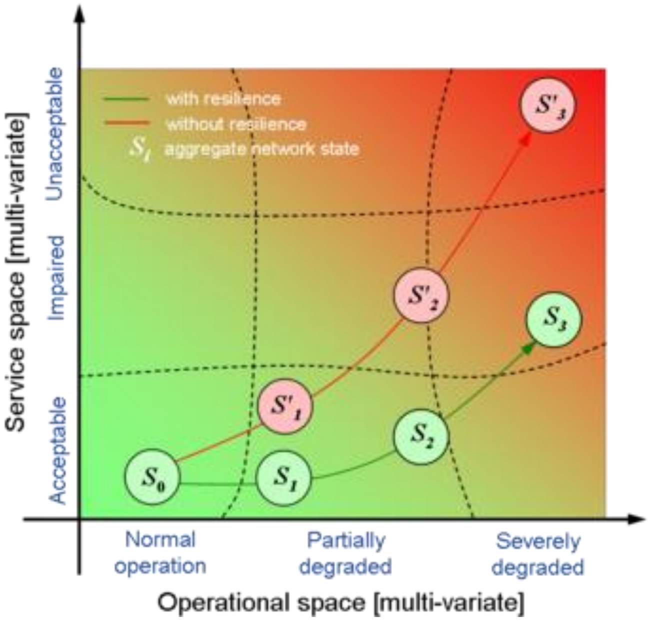
Figure 8: ResiliNets employs a two-phase strategy $\mathbf{D^2R^2}$+DR


ResiliNets
For DCN
Inner: Fault tolerance capability of products
Middle: Proactive prevention + recovery capability
Outer: Positioning + Improvement
Diagnose
Defend
Detect
Defend
Recover
Remediate
Refine
From: Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines(2010), James PG Sterbenz
Enhanced ResiliNets
For MCN
Inner kernel: high-stability architecture
Surge Tolerance
Defend
Detect
Defend
Refine
Diagnose
Recover
Extreme Bypass
Outer ring: Surge Tolerance, Extreme Bypass
Middle ring: Resilience building, proactive defense, single fault recovery, subhealth Recovery
service requirements
normal behaviour
threat and challenge models
metrics
heterogeneity
resource tradeoffs
complexity
state management
self-protection
connectivity
redundancy
diversity
multilevel
context awareness
translucency
self-organising and autonomic
adaptable
evolvable
resilience
prerequisites
tradeoffs
enablers
behaviour

Source: ResiliNet Project – James Strerbenz and Hutchison et al., 2009.

ResiliNet introduces a rigorous framework to quantify network resilience on the basis of two orthogonal dimensions as shown in Figure 4. In the figure, the Y-axis is the service space comprising three categories: acceptable, impaired, and unacceptable, and the X-axis is the operational space which comprises Normal operation, Partially degraded, and severely degraded. Thus, the two dimensions are each divided into three regions. The Resilience $\mathcal{R}$ is thus a function of state transition probability in two-dimensional state-space. The network state $\mathcal{S}$ is discrete set of operational metrics and service parameters.

Figure 9: Two dimensions when quantifying resilience



Source: ResiliNet Project – ReStrerbenz and Hutchison et al., 2009.

Topology and Challenge Modelling

An essential aspect of resilient network design is to understand how the networks behave under various challenges. If there are *n* networks for *c* challenges then a total of c × n input files would be required in simulation scripts. The methodology adopted in ResiliNets reduced the total number of input files needed to c + n input files. Topology generators are then used in the ResiliNet framework in order to gain insights into the network design, survivability analysis and node-positioning and cost optimization.

Reliability Measurement Standards

CT:TL9000 Standard

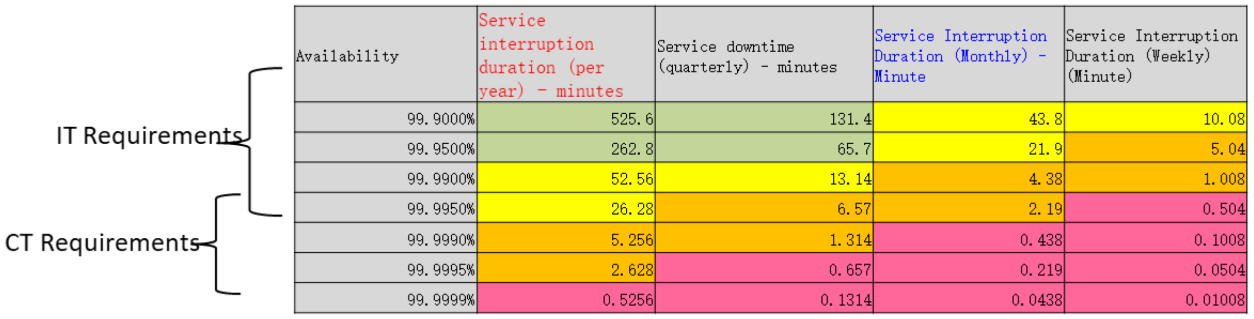
- Calculate/Measure based on the annual service interruption, update each year, and focus on the results of the current year.
- In practice, at least our company uses the interruption duration of quality accidents as the measurement of availability.
- Weighted by the number of users affected by the accident/total number

- Use the number of in-service products on the live network as the denominator to perform the weighted average.

IT Cloud: Lacks Standard

- O&M and customers are more concerned about monthly/quarterly service interruptions. (Sometimes annual availability is also mentioned)
- Currently, operators can use the interruption duration of quality accidents as the availability measurement.
- Not weighted. Some IT vendors may use the "number of users affected by the accident/total number" as weighted. The industry lacks information about this part. Generally, all faults occur in a certain area.
- Generally, only one site (region/AZ/DC) is concerned. A single incident has a great impact.

Figure 10: The objective of reliability is to minimize the service interruption time.



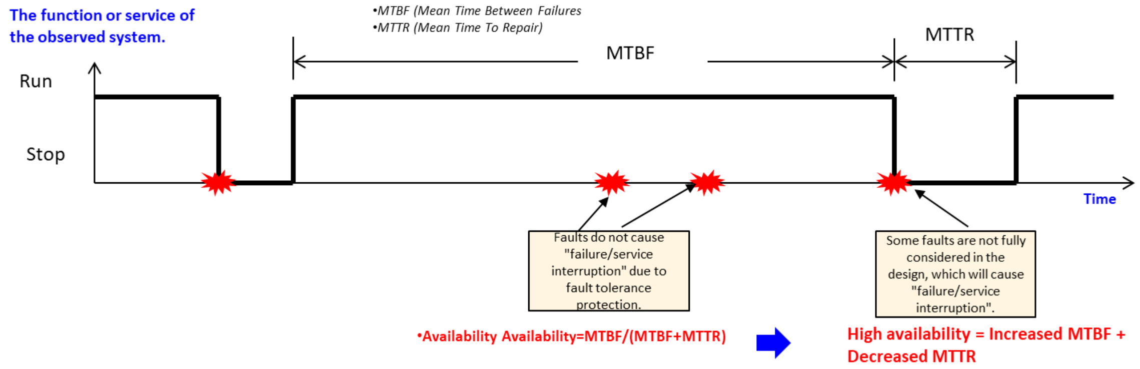
Availability	Service interruption duration (per year) - minutes	Service downtime (quarterly) - minutes	Service Interruption Duration (Monthly) - Minute	Service Interruption Duration (Weekly) (Minute)
99.9000%	525.6	131.4	43.8	10.08
99.9500%	262.8	65.7	21.9	5.04
99.9900%	52.56	13.14	4.38	1.008
99.9950%	26.28	6.57	2.19	0.504
99.9990%	5.256	1.314	0.438	0.1008
99.9995%	2.628	0.657	0.219	0.0504
99.9999%	0.5256	0.1314	0.0438	0.01008

Source: Thoth Advisory

Defining Network Faults

There are various terms that are often used interchangeably such as "error, defect, fault, failure, service interruption and outage and it is difficult to unify the various terms in the industry. In a fault-tolerant system a "fault" does not cause "failure/service interruption" due to the fault tolerant protection.

Figure 11: Faults do not cause "service interruption" in a fault tolerant system


The function or service of the observed system.
•MTBF (Mean Time Between Failures
•MTTR (Mean Time To Repair)
MTBF
MTTR
Run
Stop
Time
Faults do not cause "failure/service interruption" due to fault tolerance protection.
Some faults are not fully considered in the design, which will cause "failure/service interruption".
•Availability Availability=MTBF/(MTBF+MTTR)
High availability = Increased MTBF + Decreased MTTR

Source: Thoth Advisory

Figure 12: 70% of outages due to change management

Typical first year for a new cluster
~ 0.5 overheating (power down most machines in <5 min, ~ 1-2 days to recover
~ 1 PDU failure (~ 500-1000 machines suddenly disappear, ~ hours to be back online)
~ 1 rack-move (plenty of warning, ~500-1000 machines powered down, ~ 6 hours)
~ Network rewiring (rolling ~5% of machines down over 2-day span)
~ 20 rack failures (40-80 machines instantly disappear, 1-6 hours to recover)
~ 5 racks start acting up (40-80 machines are 50% packetless)
~ 8 network maintenance (4 might cause 30-min random connectivity losses)
~12 router reloads (takes out DNS and external links for a couple of minutes)
~ dozens of minor 30-sec blips for DNS
~ 1000 individual machine failures
~ thousands of hard drive failures

Source: Designs, Lessons and Advice from Building Large Distributed Systems Jeff Dean

When evaluating specific software objects in the network there are certain signs to look for that might indicate a warning for a fault such as:

- Oversize, over-the-limit, overload, overflow
- Redundant and lost
- Deadlock, infinite loop, leakage and damage
- Invalid status value, logic error or check error

In terms of the relationship between the object and other object in the network, these are tell-tale signs of something not working properly:

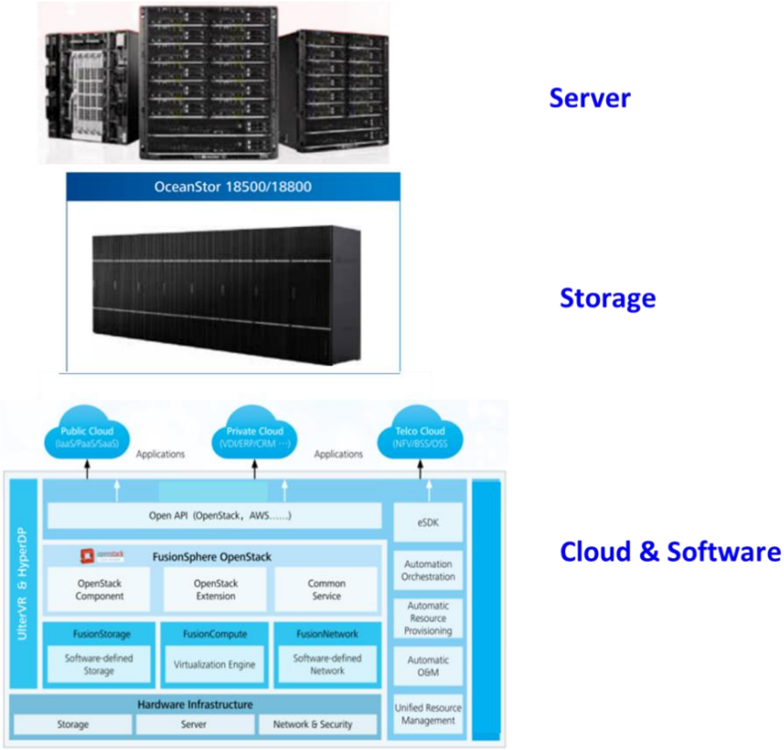
- Conflict and congestion
- Errored packets, lost packets, and disordered packets
- One-way audio, interruption, and delay
- Frequent, repeated, repeated, nested
- Mismatch, incomplete, and inconsistent
- Uneven and unbalanced

Fault Management

The fault management process involves six steps as shown in Figure 11. The three key categories of cyber physical systems in which faults can occur (detected and undetected) are:

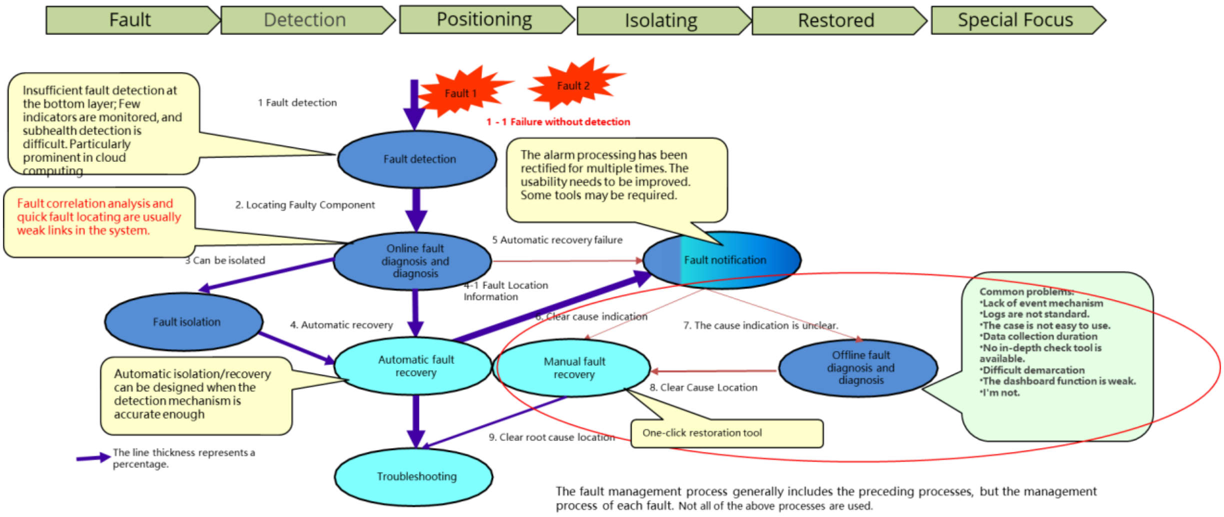
- **CPU Boards.** The key culprits are usually the CPUs, memory, HDD, cooling fans and electric power supply. Failure rates must be measured and tabulated. Troubleshooting requires comprehensive fault detection and notification capabilities for upper-layers services. Wherever feasible redundant components are included for automatic and quick recovery.
- **Storage Systems**. The main areas to focus on are the power supply, controller circuits, SSD/HDD, and I/O subsystem. Measurements that are needed are data unavailability/Data Loss, rate and duration. Troubleshooting is done as a closed system, and fault detection and recovery are all handled by the system itself. Proactive O&M is needed to eliminate potential accidents in advance.
- **Cloud and Software**. Focus should be on large-scale faults caused by network, storage, change management operations, security attacks, infrastructure faults, and service performance deterioration caused by subhealth. Measurement parameters are service health (availability, performance and capacity). Troubleshooting requires comprehensive monitoring and intelligent O&M system to enhance the detection, location, and handling of subhealth faults. Basic performance data and status alarms information needs to be provided for upper layers.

Figure 13: The three major cyber-physical systems



Source: Huawei

Figure 14: The six steps in fault management


Fault
Detection
Positioning
Isolating
Restored
Special Focus
Fault 1
Fault 2
1 - 1 Failure without detection
1 Fault detection
Insufficient fault detection at the bottom layer; Few indicators are monitored, and subhealth detection is difficult. Particularly prominent in cloud computing
Fault detection
The alarm processing has been rectified for multiple times. The usability needs to be improved. Some tools may be required.
2. Locating Faulty Component
Fault correlation analysis and quick fault locating are usually weak links in the system.
Online fault diagnosis and diagnosis
5 Automatic recovery failure
Fault notification
3 Can be isolated
4-1 Fault Location Information
Fault isolation
4. Automatic recovery
6. Clear cause indication
7. The cause indication is unclear.
Common problems:
•Lack of event mechanism
•Logs are not standard.
•The case is not easy to use.
•Data collection duration
•No in-depth check tool is available.
•Difficult demarcation
•The dashboard function is weak.
•I'm not.
Automatic fault recovery
Manual fault recovery
Offline fault diagnosis and diagnosis
8. Clear Cause Location
Automatic isolation/recovery can be designed when the detection mechanism is accurate enough
One-click restoration tool
9. Clear root cause location
Troubleshooting
The line thickness represents a percentage.

The fault management process generally includes the preceding processes, but the management process of each fault. Not all of the above processes are used.

Source: Thoth Advisory

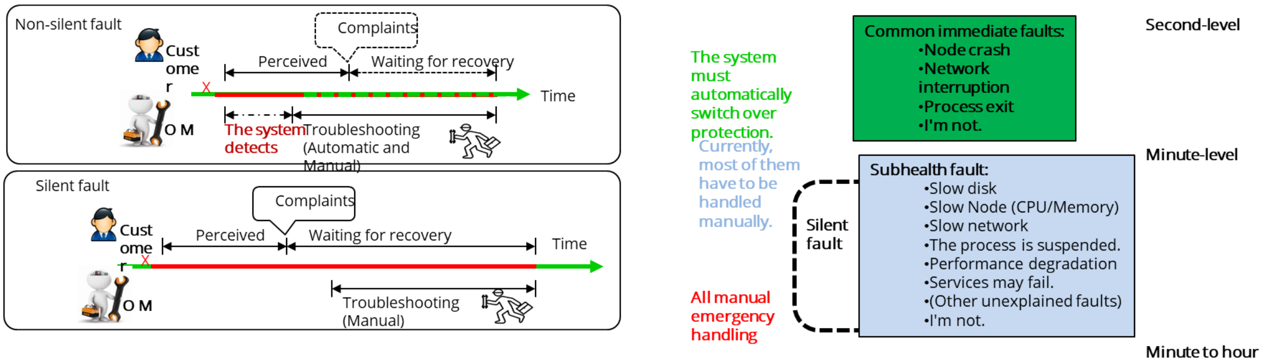
Fault Detection

Checklist for Fault Management

1. Are there any symptoms that could predict the failure? For example, observe some performance indicators? What is the change rule? Can I find it through the health check? Is real-time monitoring necessary? Is there a significant impact on performance during real-time monitoring?
2. Is there a detection in case of failure? How fast do you want to test? Check whether services are affected or data is inconsistent during the detection. Will there be misjudgment/missing judgment? Can I detect traffic congestion? Is there flow control? Will VIP customers be guaranteed?
3. Which of the following operations are high-risk operations? Are there any faults during high-risk operations? What do I need to pay attention to? How do you tell the customer this?
4. Does the problem affect user services? What is the impact on upper-layer services? Is the impact wide? Is isolation required? Is it fault-prone? If a large-scale fault occurs on the entire system, does it need to be automatically rectified immediately?
5. Is the location accurate? Do you want to restore automatically? Is automatic recovery or alarm reporting first? Automatic recovery has not worked. Do you want to try again? What can I do if I retry multiple times but no effect occurs? Does auto-recovery affect performance?
6. Are there any emergency measures for service interruption due to major faults? Is it easy to operate according to the information?
7. What if management configuration data is lost? Do you want a scheduled backup? Can I back up immediately when a change occurs? How many backups are appropriate? What if the directory is full of backups? Where to back up?
8. Logged? Is the log written in a standard manner? Is the log easy to use? Will the directory be fully occupied when logs are added? Do you want to spit out the alarm? How to write the alarm help? Are many alarms generated at a time?
9. Is the hardware fault easy to replace? Is the fault located to the FRU? Will services be interrupted during the replacement? What is the strategy for onsite spare parts? What is the parts replacement rate?
10. Do you want to back up user data? Is the user backing up the data themselves, or is the system all arranged?
11. Is fault isolation adequate? Does the fault or performance deterioration of a single node affect the entire system? Can all faults or performance degradation be detected by observing business metrics?

Figure 15: Fault Detection

Fault | Detection | Positioning | Isolating | Restored | Special Focus



Source: Thoth Advisory

Common methods to detect faults:

1. Value range, such as temperature, voltage, slot ID, and MAC address

2. Data correctness check, such as parity check, checksum calculation, and CRC check

3. Comparison check: generally refers to the comparison of redundant system output values; For example, check the consistency of the configuration data between the foreground and background.

4. Check the time, such as the heartbeat detection.

Precautions:

1. **Correctness:** Check for multiple times to avoid false detection. The detection method must be accurate and all scenarios must be considered to avoid false detection.

2. **Security:** Detection behavior cannot occupy too many system resources to avoid affecting performance. The detection behavior should not affect the running of ongoing services.

3. **Relevance:** The period for detecting lower-layer faults must be shorter than the period for detecting upper-layer faults. Design a method for detecting the consistency between the detection path and service path based on different scenarios.

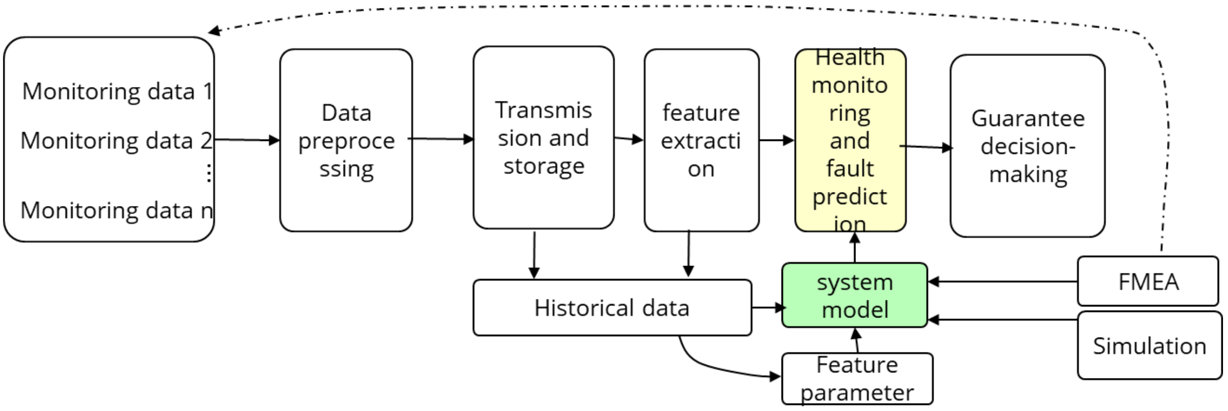
4. **Criticality:** Systematically design the faults that need to be detected and ensure that the detection mechanism (such as heartbeat) has a high priority.

5. **Detection difficulty:** 1. Subhealth detection; 2. Reduce silent faults. 3. Service availability detection;

Figure 18 shows illustrates the steps that are needed when designing the network to prevent service interruption. The time series probability trend, artificial neural network, and cluster analysis methods can be used to establish the relationship model between abnormal phenomena and health status and predict faults. For example, Artificial neural network (ANN) can imitate the ability of continuous nonlinear function, and can learn from samples, master the system law, has strong self-adaptive learning ability. However, neural network training requires a large number of data samples, and there are disadvantages such as slow convergence speed and difficult determination of local abnormal points.

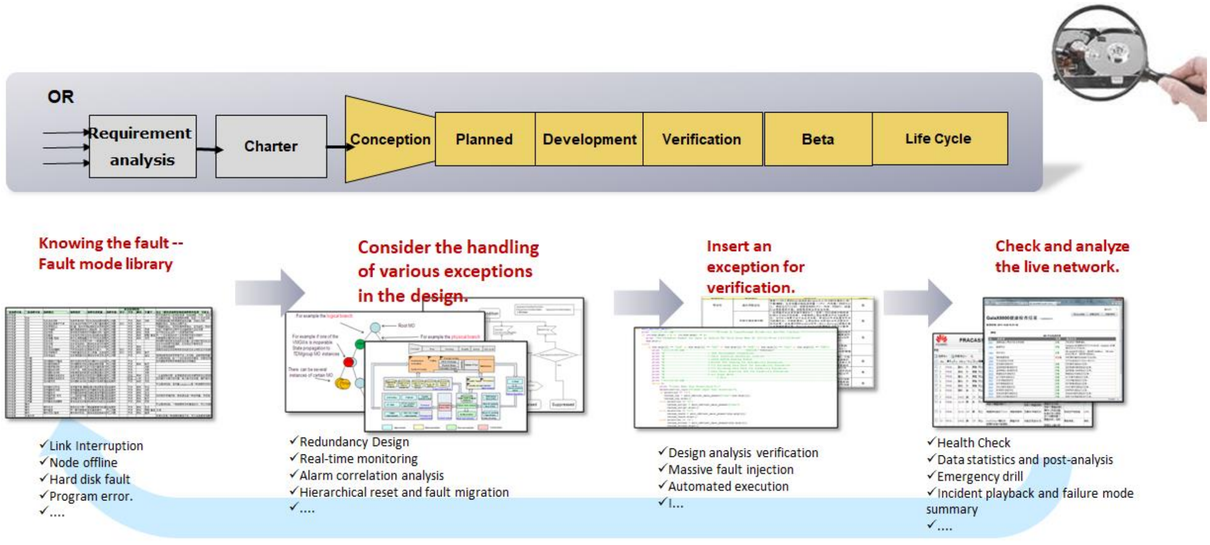
Time series prediction method is to arrange the historical data of the prediction object according to a certain time interval, form a statistical series with time change, establish the corresponding data change model with time, and extrapolate the model to the future for prediction. This approach is effective if the past development model continues into the future.

Figure 16: The highest level of fault detection is fault prediction



Source: Thoth Advisory

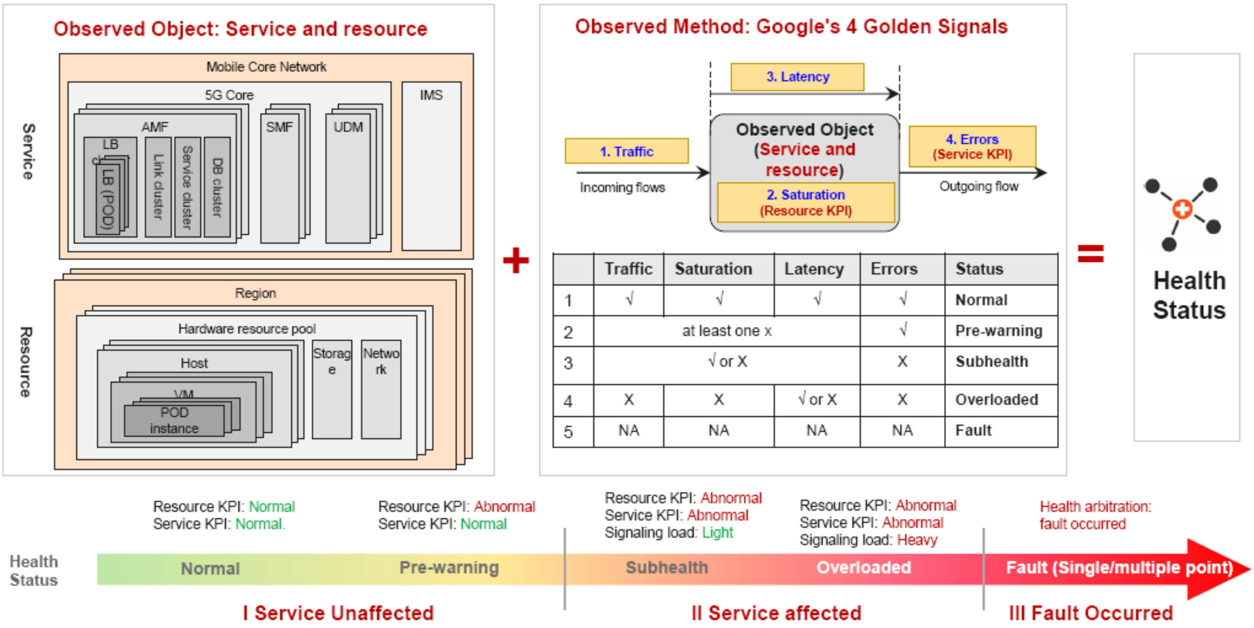
Figure 17: Fault-based design prevents service interruption caused by these faults in the future.


OR
Requirement analysis
Charter
Conception
Planned
Development
Verification
Beta
Life Cycle
Knowing the fault -- Fault mode library
Consider the handling of various exceptions in the design.
Insert an exception for verification.
Check and analyze the live network.
✓Link Interruption
✓Node offline
✓Hard disk fault
✓Program error.
✓....
✓Redundancy Design
✓Real-time monitoring
✓Alarm correlation analysis
✓Hierarchical reset and fault migration
✓....
✓Design analysis verification
✓Massive fault injection
✓Automated execution
✓I...
✓Health Check
✓Data statistics and post-analysis
✓Emergency drill
✓Incident playback and failure mode summary
✓....

Source: Thoth Advisory

A typical Google cluster has about 1000 servers. Google SRE has found that roughly 70% of outages are due to change management events. The right side of Figure 20 shows Google's four Golden signal metrics: Traffic, Saturation, Latency, and Errors and the strategy for classifying the health of the network. For example, if Traffic/Saturation/Latency are normal but Errors are occurring then this is a *Pre-warning*. If one or more of Traffic/Saturation/Latency are showing problems but still reporting metrics and errors are occurring then is treated as *Subhealth*. A *fault* then is interpreted when all the metrics become unavailable.

Figure 18: Definition of Health Status and Google's Four Golden KPI Thresholds



	Traffic	Saturation	Latency	Errors	Status
1	√	√	√	√	Normal
2	at least one x			√	Pre-warning
3	√ or X			X	Subhealth
4	X	X	√ or X	X	Overloaded
5	NA	NA	NA	NA	Fault

Source: Thoth Advisory

Subhealth Detection in Storage Systems

Hardware failure of a single component is a basic problem for all distributed systems. The solution is to automatically detect and recover the hardware failure without affecting services and customer experience. The defect of the distributed software architecture is that as the performance of a single node decreases then the performance of the entire system to deteriorate.

- On the one hand, the resource management of the single-node survival environment ensures the health of the node.
- Even if a single node is faulty, the system performance should not deteriorate.

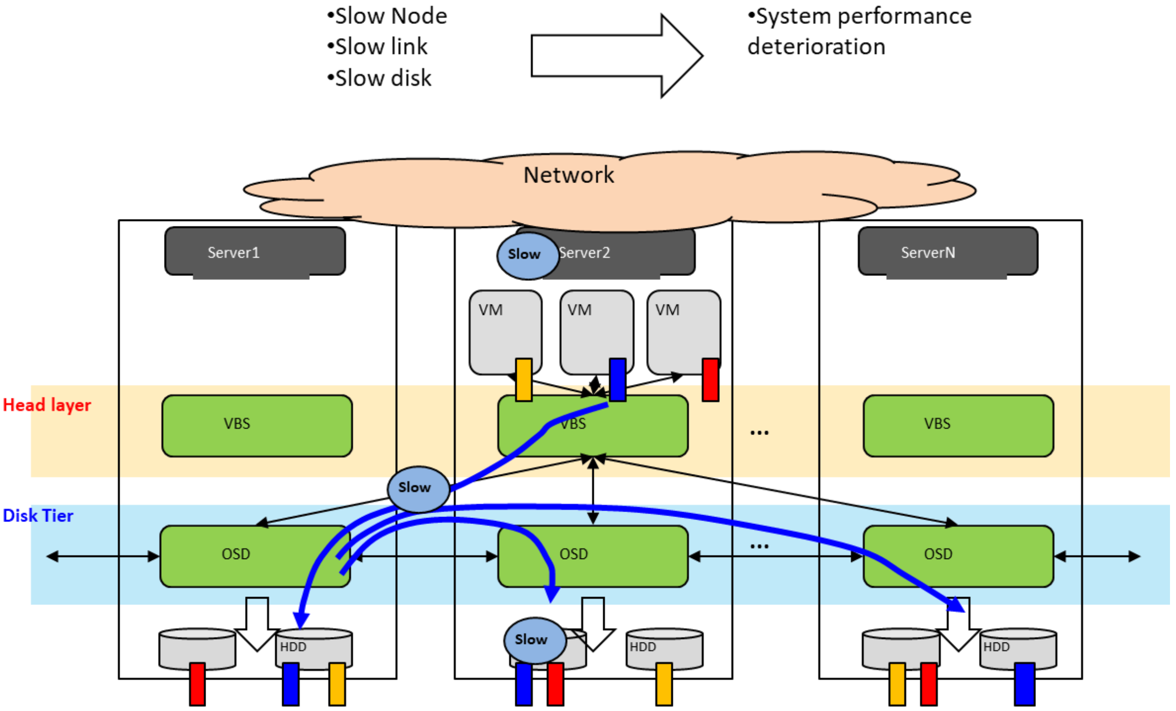
Figure 21 shows an actual example of subhealth detection. Due to a bug in the new UVP software version, *ulimit* fails to modify the number of FDs. The maximum number of FDs in the eMDC network module is 1024. As a result, the OSD (Object-based Storage Device) fails to establish a link with the eMDC and cannot be started. The MDC (Multi-tenant Database Container), OSD, VBS (Visual Basic

Script), and relational databases are deployed on the same node. Many small volumes are created. OSD treats individual data set as ab object with its own metadata and identifiers.

The LVM (Logical Volume Management) is used to create a large volume for the Oracle database. The relational database has many partitions on the large volume, including the /home directory. This problem occurs when a large file is copied to the /home directory. (XFS used by the LVM, EXT3 used by the /home partition, and sync mode. XFS is a high-performance journaling file system) The node is configured with 250 GB memory, and the remaining memory is about 1 GB for a long time. As a result, there are too many memory fragments. When the TCP protocol stack does not apply for memory when sending packets, the memory is sorted. The kernel preempts the CPU of service processes in a short period of time when the memory is sorted. As a result, the OSD heartbeat packet may not be sent in time. In addition, memory sorting is performed when the TCP protocol stack sends packets, resulting in a long I/O delay. Heartbeat packets may be blocked in the NIC sending queue due to service packet congestion. As a result, heartbeat packets cannot be sent in time and the MDC reports the abnormality.

As long as HDDs are being used in the telco network there will be issues with the quality and speed of disks which can affected by many factors. In addition, applications have different requirements on disks. For example, online focuses on latency, offline focuses on throughput, and more subdivided indicators focus on service life. Therefore, scoring rules are dynamically adjusted. You can specify the weight and calculation rules of each indicator for each application.

Figure 19: Example of subhealth detection



Note: OSD = Object-based Storage Device, VBS = Visual Basic Script

Source: Thoth Advisory

The number of times that a thread of the Kernel is in the D state determines the disk speed. If the thread is in the D state for a long time, the disk must be deemed to be faulty. The disk status changes as follows:

- When a disk enters the Slow or WARNING state, online applications do not use the disk because the disk may be damaged, which greatly increases the latency. However, offline applications can continue to use the disk.
- If the disk enters the ERROR state, it may be damaged immediately. In this case, you must take the disk offline immediately. In this case, the system identifies the disk through the disk state machine and sets the disk to the Shutdown state. That is, the disk cannot be read or written and starts to back up data.
- After the system confirms that the data backup is complete (data security), the disk enters the REPAIR state. In the Repair state, the system will report for repair through the O&M maintenance system.
- A supplier might replace disks two to three times a week based on the status of the repair request system. After the replacement, the system will automatically detect that a new disk goes online and enter the NEWBIE state. For disks in the NEWBIE state, the system automatically partitions, formats, and pre-mounts the disks.
- After the pre-mounting is complete, the system enters the TEST state. After basic check and small benchmark test, Hua Tuo decides whether to continue the repair or go online.

- The disks that pass the test will be mounted to the actual mount point for use. At the same time, the disk enters the GOOD state.
- All disk data is mined and analyzed by the analyzer in the ODPS. After the analysis is complete, the disk detection accuracy is improved by adjusting rules, indicators, and parameters. This is a long-term process.

Setting thresholds is an important feature in fault detection (See Figure 22 below).

Figure 20: Setting suitable threshold conditions

Threshold	Description
3.3 Rule (default)	When 3 consecutive threshold events occurring within 3 consecutive sampling intervals are observed.
4-in-1 Rule	When 4 threshold events are observed withing a 24-hour period
4 Decreasing	When 4 threshold events are observed over decreasing sampling intervals
Custom	Allows a user to modify the permutations of the 3.3 and the 4-in-1rules.

Source: Thoth Advisory

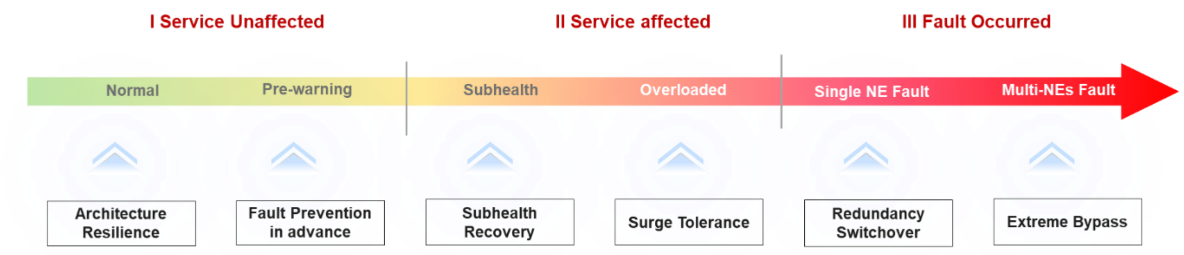
If the sampling interval is 15 minutes then it will take 45 minutes in order to identify a slow disk. Two thresholds "Unacceptable to Customer Service" and "Unacceptable to Customer Service" use the default values (such as with NetApp) in the early stage. Later, the thresholds can be adjusted based on the data provided by the live network and test.

The short detection period is set to 5 minutes. The algorithm determines the status of the 5-minute detection period. The large detection period is set to 30 minutes. If an exception is reported within 30 minutes, the alarm is reported to the control node. If the exception persists, the algorithm continuously reports the exception (the control node filters the exception by itself). For slow, the algorithm generates logs every 30 minutes.

Developing a 3-6 Fault Classification System

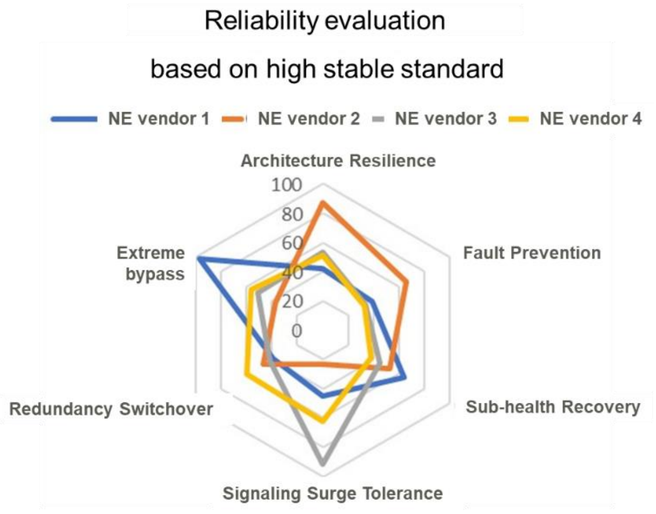
Figure 23 summarizes the proposed 3-6 Fault Classification System and Figure 24 illustrates how the 6 dimensions can be viewed in a Spider Chart format to gain insights into the overall stability of the network.

Figure 21: Three-Six (3-6) Classification System



Source: Thoth Advisory

Figure 22: Example of Reliability Evaluation using 6 Dimensions



Source: Thoth Advisory

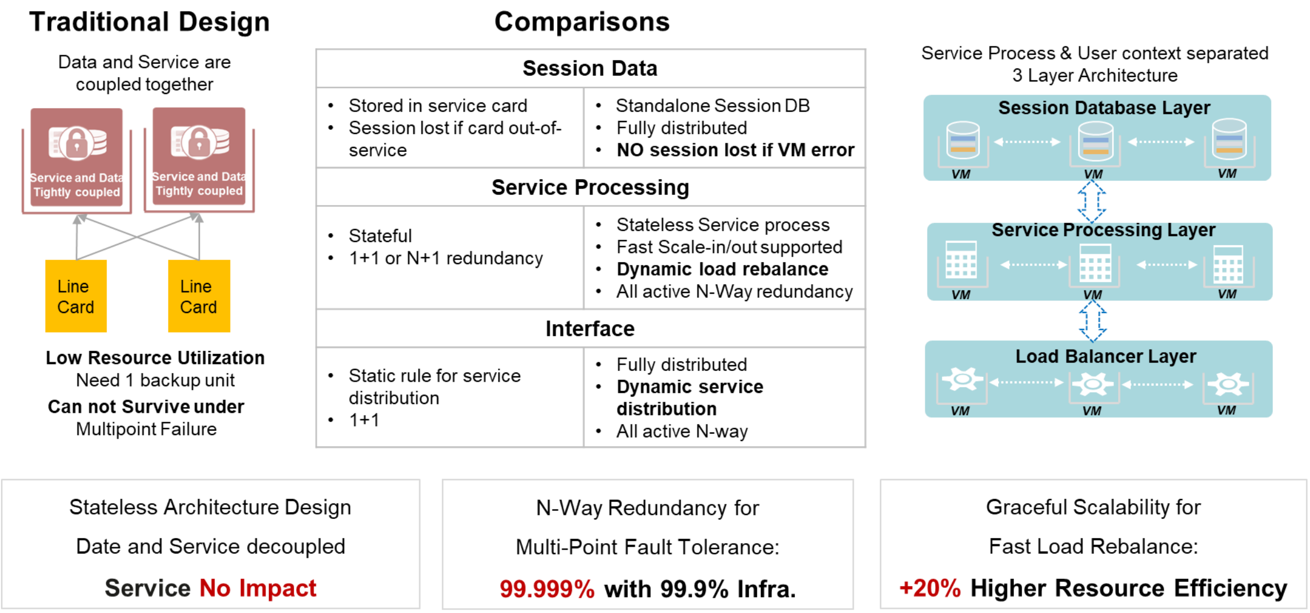
Proposing 6 Types of Network Actions

Architecture resilience

In traditional design data and services are coupled together with redundancy provided by cross links but this approach is vulnerable and cannot survive multi-pint failure. N-Way redundancy organized as three layers ensures redundancy at three levels: session data, service processing and the link interfaces.

Anti-multipoint failure of infrastructure. The system should support separation of programs and data, preventing services from being affected when multiple points of fault occur.

Figure 23: N-Way Software Architecture Redesign for Reliability Improvement



Traditional Design

Data and Service are coupled together

Service and Data Tightly coupled
Service and Data Tightly coupled
Line Card
Line Card

Low Resource Utilization
Need 1 backup unit
Can not Survive under
Multipoint Failure

Comparisons

Session Data	
• Stored in service card • Session lost if card out-of-service	• Standalone Session DB • Fully distributed • **NO session lost if VM error**
Service Processing	
• Stateful • 1+1 or N+1 redundancy	• Stateless Service process • Fast Scale-in/out supported • **Dynamic load rebalance** • All active N-Way redundancy
Interface	
• Static rule for service distribution • 1+1	• Fully distributed • **Dynamic service distribution** • All active N-way

Service Process & User context separated
3 Layer Architecture

Stateless Architecture Design
Date and Service decoupled
Service No Impact

N-Way Redundancy for
Multi-Point Fault Tolerance:
99.999% with 99.9% Infra.

Graceful Scalability for
Fast Load Rebalance:
+20% Higher Resource Efficiency

Source: Thoth Advisory

Fault Prevention via a Proactive defense

20 core network global incidents have occurred in the past four years affecting 200 million users. 39% of these were attributed to DR failures such as:

- Manual evaluation is time consuming, inadequate , and prone to secondary disasters.
- Flow control parameters are set based on experience and are inadequate

- Operations are directly performed on the network without simulation , leaving unknown
- The DR process is invisible , and problems cannot be detected promptly.

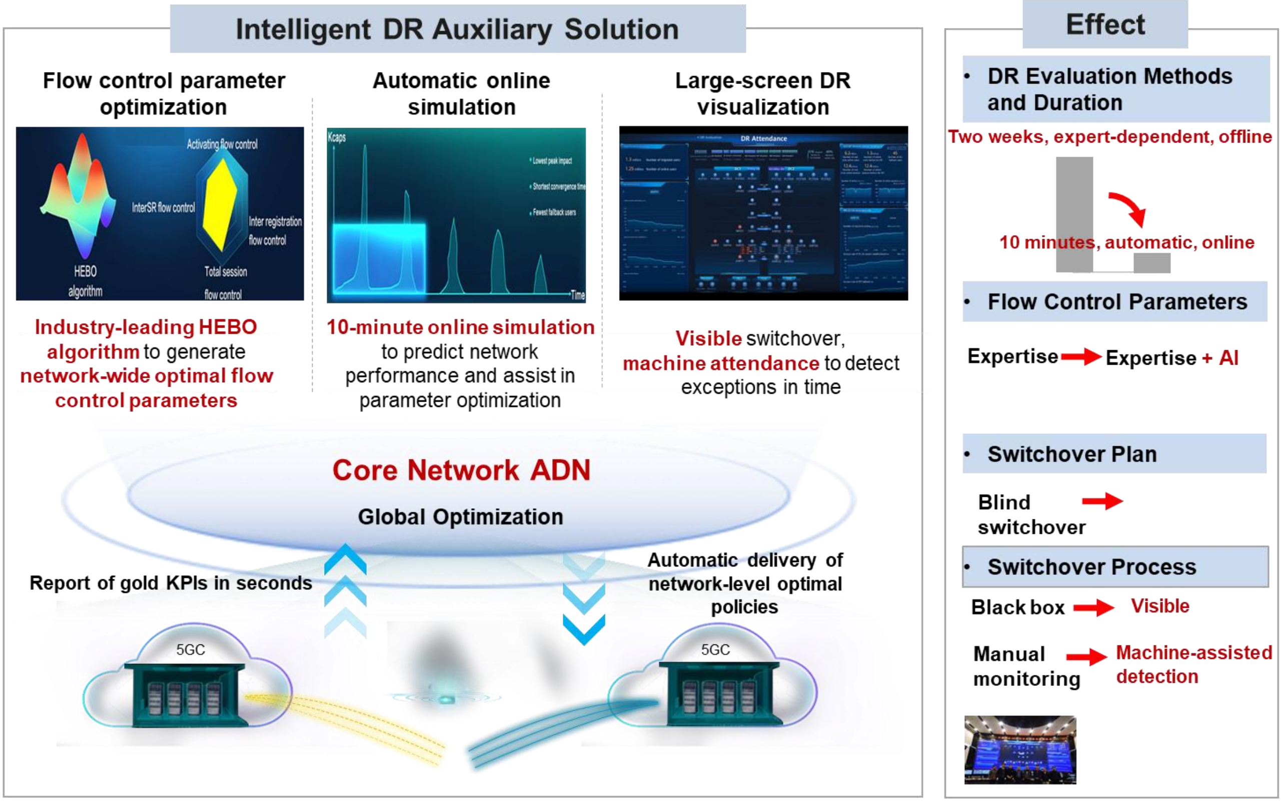
Figure 26 shows an example of an intelligent DR solution that uses a new algorithm HEBO to generate network-wide optimal flow control parameters.

Resource exception prediction. The system should support resource leakage and abnormal usage are detected 24 hours before service interruption.

Mis-operation prevention and control. The system should support protective mechanism for mis-operations and prevent service loss.

Service exception prediction. The system should support predicting service traffic and function exceptions to prevent *fault spread*.

Figure 24: Intelligent Data Recovery safeguards DR switchover



Intelligent DR Auxiliary Solution

Flow control parameter optimization

Industry-leading HEBO algorithm to generate network-wide optimal flow control parameters

Automatic online simulation

10-minute online simulation to predict network performance and assist in parameter optimization

Large-screen DR visualization

Visible switchover, machine attendance to detect exceptions in time

Core Network ADN

Global Optimization

Report of gold KPIs in seconds

Automatic delivery of network-level optimal policies

Effect

- DR Evaluation Methods and Duration

Two weeks, expert-dependent, offline

10 minutes, automatic, online

- Flow Control Parameters

Expertise → Expertise + AI

- Switchover Plan

Blind switchover →

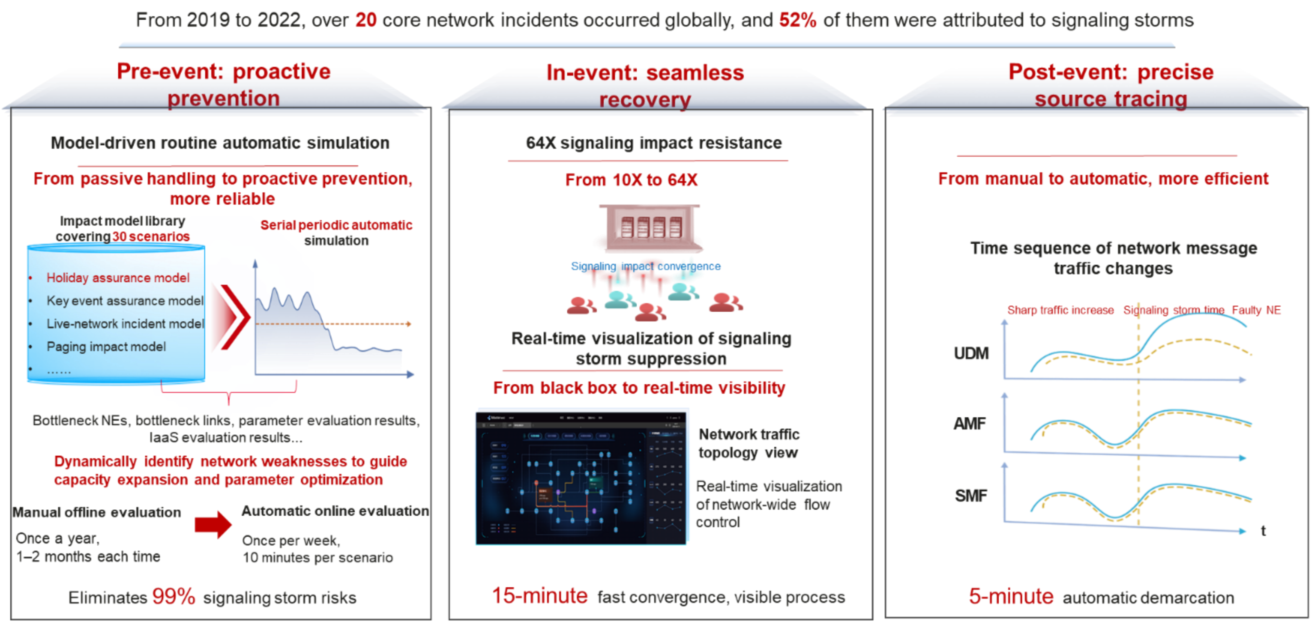
- Switchover Process

Black box → Visible

Manual monitoring → Machine-assisted detection

Source: Thoth Advisory

Figure 25: Intelligent Signalling Storm Prevention



From 2019 to 2022, over 20 core network incidents occurred globally, and 52% of them were attributed to signaling storms

Pre-event: proactive prevention

Model-driven routine automatic simulation

From passive handling to proactive prevention, more reliable

Impact model library covering 30 scenarios

- Holiday assurance model
- Key event assurance model
- Live-network incident model
- Paging impact model
- ……

Serial periodic automatic simulation

Bottleneck NEs, bottleneck links, parameter evaluation results, IaaS evaluation results...

Dynamically identify network weaknesses to guide capacity expansion and parameter optimization

Manual offline evaluation

Once a year,
1–2 months each time

Automatic online evaluation

Once per week,
10 minutes per scenario

Eliminates 99% signaling storm risks

In-event: seamless recovery

64X signaling impact resistance

From 10X to 64X

Real-time visualization of signaling storm suppression

From black box to real-time visibility

Network traffic topology view

Real-time visualization of network-wide flow control

15-minute fast convergence, visible process

Post-event: precise source tracing

From manual to automatic, more efficient

Time sequence of network message traffic changes

5-minute automatic demarcation

Source: Thoth Advisory

Subhealth recovery

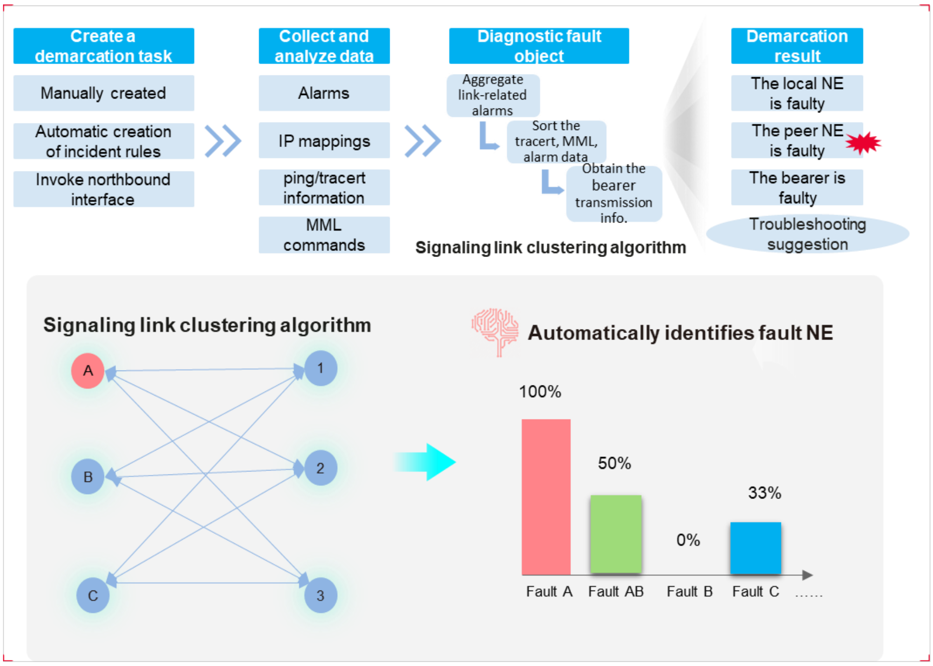
Module redundancy. The system should support redundancy switching capability (in the modules and NEs) in the hardware.

Automatic recovery. Moreover, the system should support 5-minute diagnosis, 15-minute recovery, and zero service interruption.

Individual Fault Recovery. The system should individual fault recovery which can be realized by performing cross-NE and Cross-Layer diagnosis and restoration. Group fault awareness and session failover are also very useful features to have.

More than 90% of link issues are caused by peer-end operations or intermediate transmission operations faults. The feature can automatically demarcate the fault to the local or peer NE and the bearer channel. Signaling link faults occur frequently and handling such faults is time-consuming. The solution is to implement intelligent fault diagnosis as shown in the Figure 28 below.

Figure 26: Intelligent fault diagnosis via fast demarcation of link faults


Create a demarcation task
Manually created
Automatic creation of incident rules
Invoke northbound interface
Collect and analyze data
Alarms
IP mappings
ping/tracert information
MML commands
Diagnostic fault object
Aggregate link-related alarms
Sort the tracert, MML, alarm data
Obtain the bearer transmission info.
Signaling link clustering algorithm
Demarcation result
The local NE is faulty
The peer NE is faulty
The bearer is faulty
Troubleshooting suggestion
Signaling link clustering algorithm
A
B
C
1
2
3
Automatically identifies fault NE
100%
50%
0%
33%
Fault A
Fault AB
Fault B
Fault C
……

Source: Thoth Advisory

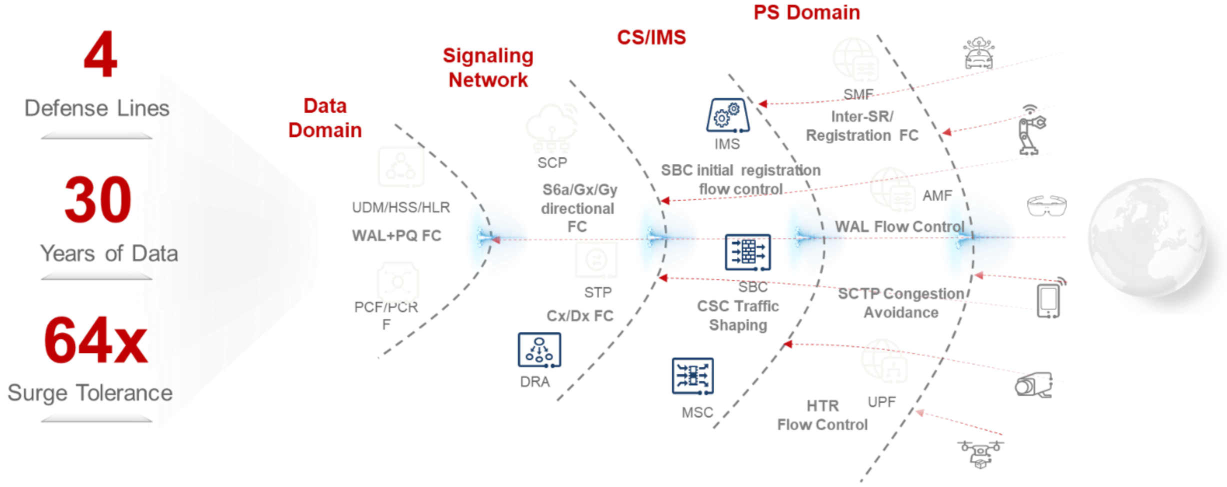
Surge Tolerance

Robust self-protection. The system should support 64 fold signaling overload control capability and surge convergence within 30 minutes. This can be done with overload control: smoothing burst signaling, priority-based package discard, and backpressure flow control.

Users can be back online quickly. The system should support active users going online within 5 minutes. This can be realized with intelligent camping, HTR flow control and SCTP (Stream Control Transmission Protocol) congestion control.

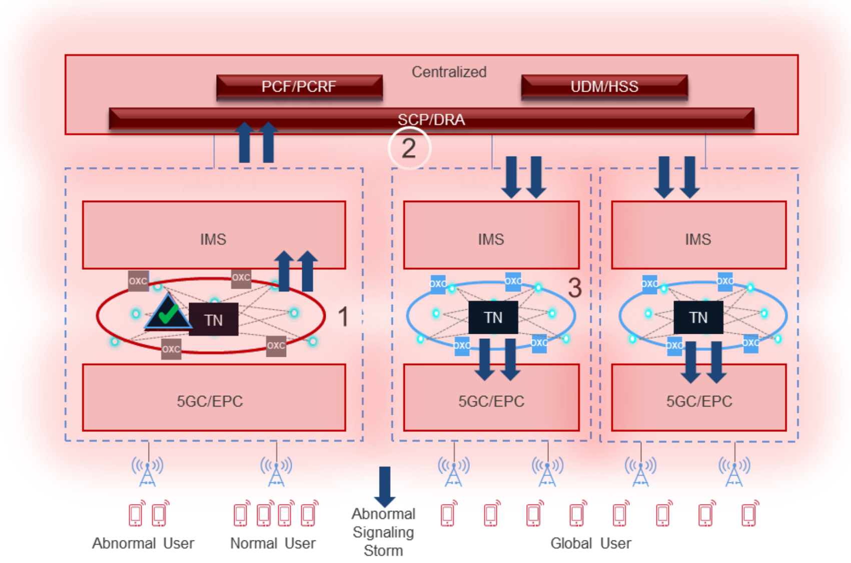
Moreover, the system should support signaling storm convergence within 5 minutes and real-time online enablement of active users.

Figure 27: 4-level signalling storm defense to prevent network outage



Source: Thoth Advisory

Figure 28: Automatic configuration of HTR flow control parameters to prevent downstream NE overload in 5G Core



1 TN failure cause IMS signaling Storms; IMS outage

2 Fault spread to signaling network; STP/DRA outage

3 Fault spread to other area; Other IMS and 5GC is hard to reach the UDM, and fail with service.

Source: Thoth Advisory

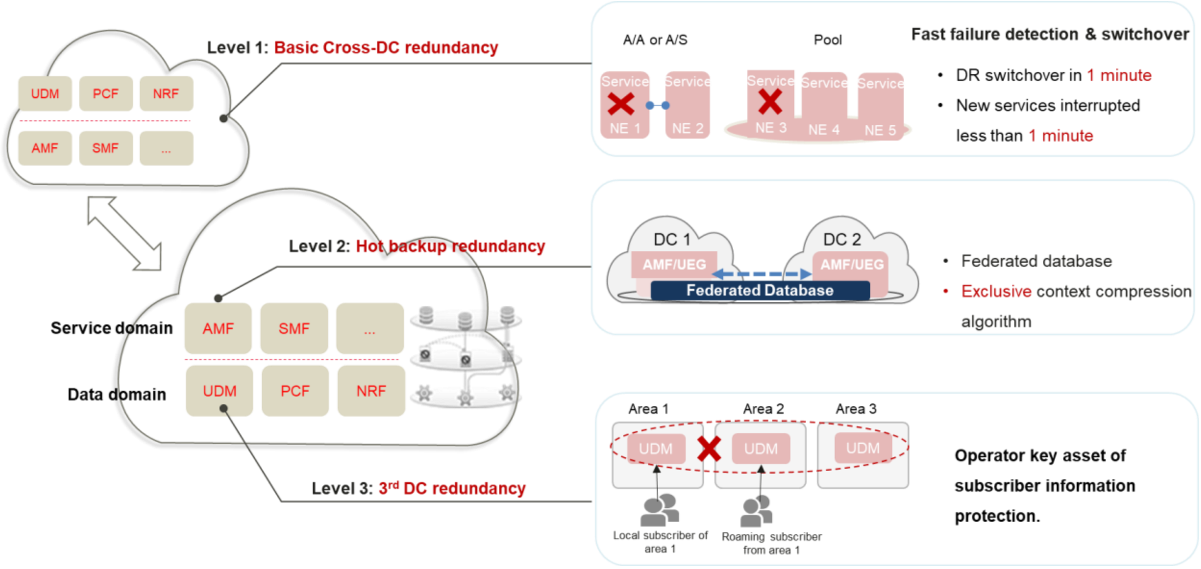
Single Fault Troubleshooting (redundancy switchover)

Basic Disaster Recovery Capability. The system should support services be recovered within 1 minute when a session related NE is faulty. This can be remedied with basic disaster recovery technology defined in 3GPP. The system should support active/standby protection for data management NEs. The system should also support disaster recovery drill.

Hot Backup Redundancy. The system should support session-related NE faults do not interrupt running services and new service request can be taken over within 1 minute and multi-active redundancy and third-DC deployment for data-related NEs.

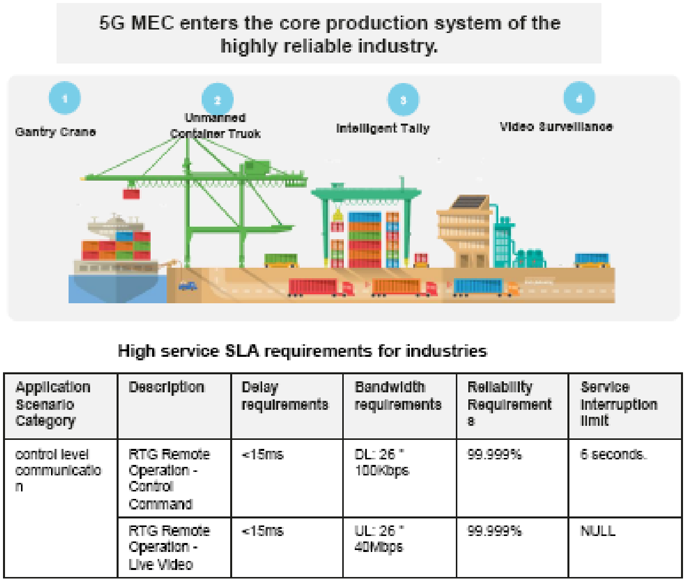
Gray Disaster Recovery. The system should support session related NE faults not interrupt running services and new service request can be taken over within 30 seconds and smaller cell N Way redundancy for data related NEs.

Figure 29: Three-level reliability design, implementing service restoration upon NVF faults


Level 1: Basic Cross-DC redundancy
UDM
PCF
NRF
AMF
SMF
...
A/A or A/S
Pool
Service
NE 1
NE 2
NE 3
NE 4
NE 5
Fast failure detection & switchover
DR switchover in 1 minute
New services interrupted less than 1 minute
Level 2: Hot backup redundancy
Service domain
Data domain
DC 1
DC 2
AMF/UEG
Federated Database
Federated database
Exclusive context compression algorithm
Level 3: 3rd DC redundancy
Area 1
Area 2
Area 3
Local subscriber of area 1
Roaming subscriber from area 1
Operator key asset of subscriber information protection.

Source: Thoth Advisory

Figure 30: Hot Backup to support 5G industrial deployments



5G MEC enters the core production system of the highly reliable industry.

High service SLA requirements for industries

Application Scenario Category	Description	Delay requirements	Bandwidth requirements	Reliability Requirements	Service interruption limit
control level communication	RTG Remote Operation - Control Command	<15ms	DL: 26 * 100Kbps	99.999%	6 seconds.
	RTG Remote Operation - Live Video	<15ms	UL: 26 * 40Mbps	99.999%	NULL

Source: Thoth Advisory

Extreme/Fault bypass

Infrastructure bypass. The system should support service inertial running in case of underlay infrastructure faults. Solutions for this include voice fallback to Circuit Switched, Packet Switch fallback to 4G, third party cold backup of UDM data, OMU bypass.

NE fault bypass. The system should support data NE fault bypass and service NE bypass and data plane inertial running in case of control plane fault. Solutions for this include UDM/PCf/NRF bypass, OCS/CHF bypass, and storage bypass.

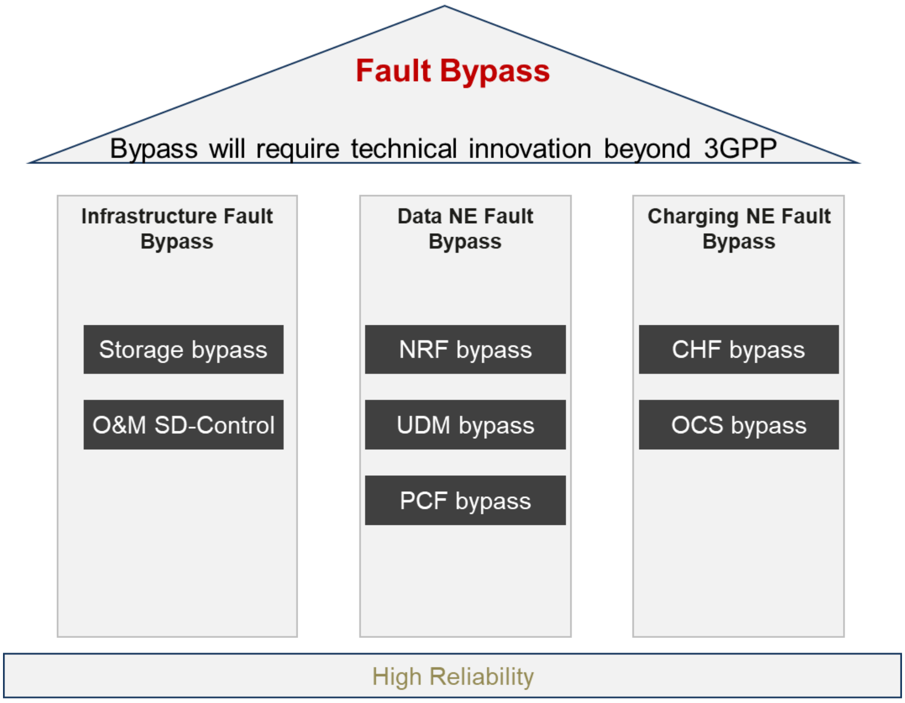
KPI Deterioration Bypass. The system should support fast handover or recovery after KPI decreases.

Figure 33 shows a generalization of different bypass innovations that vendors could implement. These are not specifically called out in the 3GPP specifications so vendors will need to provide innovation in the 5G Core to realize different types of bypass in the event of a fault.

Figure 34 illustrates bypass in the case of storage system reliability issues such as power interruption causing disk array pool fault or disk(s) are experiencing subhealth characteristics or a link between the router and the storage system suffers a problem. One way to implement a bypass in the aforementioned scenarios is to move the executable file from disk to memory; this would of course require that memory is reserved in advance.

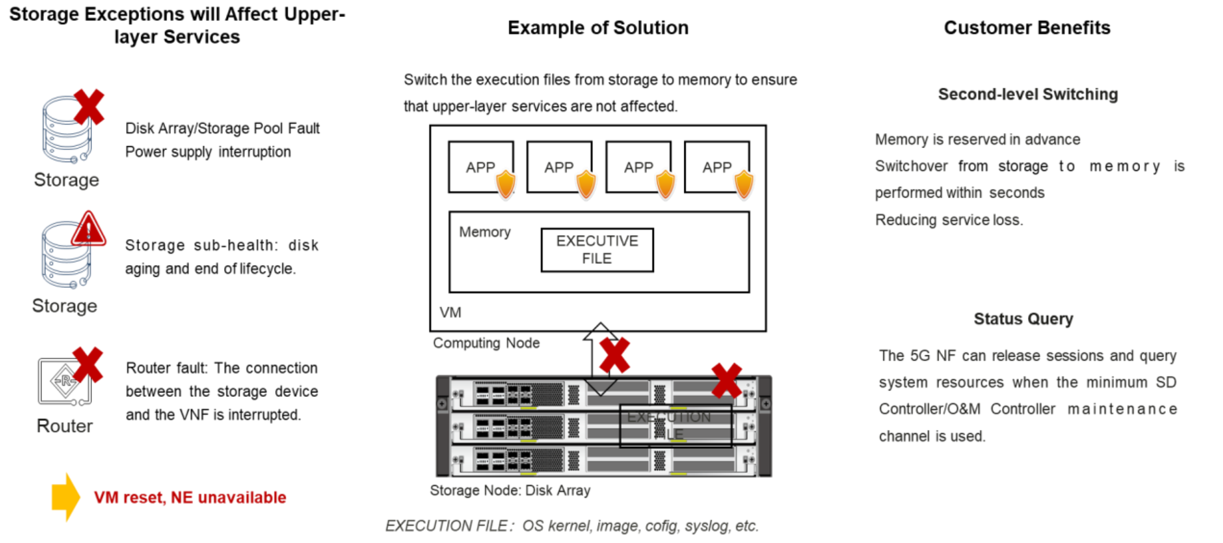
Figure 35 shows the scenario where both UDMs (Unified Data Management) are faulty. In this case the AMF is routed to the UPF (User Plane Function) using minimum subscription data configured on itself, instead of interworking with a UDM. The IMS (IP Multi-media System) will implement re-registration of the calling party without authentication, and the addressing capability of the called S-CSCF (Serving – Call Session Control Function) is enhanced. The S-CSCF is the primary node in the IMS responsible for session control. Under normal operation, subscribers will be allocated a S-CSCF for the duration of their IMS registration in order to facilitate routing of SIP messages as part of service establishment procedures.

Figure 31: Fault bypass require technical innovation above the 3GPP specifications



Source: Thoth Advisory

Figure 32: Storage Bypass to ensure service continuity



Storage Exceptions will Affect Upper-layer Services

Disk Array/Storage Pool Fault
Power supply interruption

Storage sub-health: disk aging and end of lifecycle.

Router fault: The connection between the storage device and the VNF is interrupted.

VM reset, NE unavailable

Example of Solution

Switch the execution files from storage to memory to ensure that upper-layer services are not affected.

EXECUTION FILE： OS kernel, image, cofig, syslog, etc.

Customer Benefits

Second-level Switching

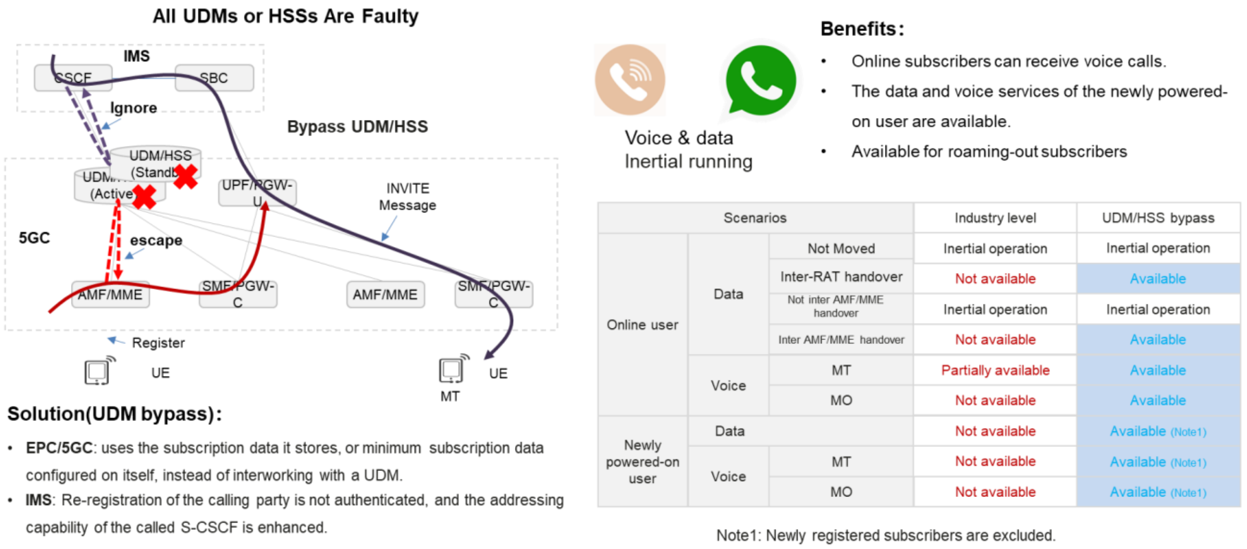
Memory is reserved in advance
Switchover from storage to memory is performed within seconds
Reducing service loss.

Status Query

The 5G NF can release sessions and query system resources when the minimum SD Controller/O&M Controller maintenance channel is used.

Source: Thoth Advisory

Figure 33: UDM/HSS Bypass for data and voice service continuity



All UDMs or HSSs Are Faulty

Solution(UDM bypass):

- **EPC/5GC**: uses the subscription data it stores, or minimum subscription data configured on itself, instead of interworking with a UDM.
- **IMS**: Re-registration of the calling party is not authenticated, and the addressing capability of the called S-CSCF is enhanced.

Benefits:

- Online subscribers can receive voice calls.
- The data and voice services of the newly powered-on user are available.
- Available for roaming-out subscribers

Scenarios			Industry level	UDM/HSS bypass
Online user	Data	Not Moved	Inertial operation	Inertial operation
		Inter-RAT handover	Not available	Available
		Not inter AMF/MME handover	Inertial operation	Inertial operation
		Inter AMF/MME handover	Not available	Available
	Voice	MT	Partially available	Available
		MO	Not available	Available
Newly powered-on user	Data		Not available	Available (Note1)
	Voice	MT	Not available	Available (Note1)
		MO	Not available	Available (Note1)

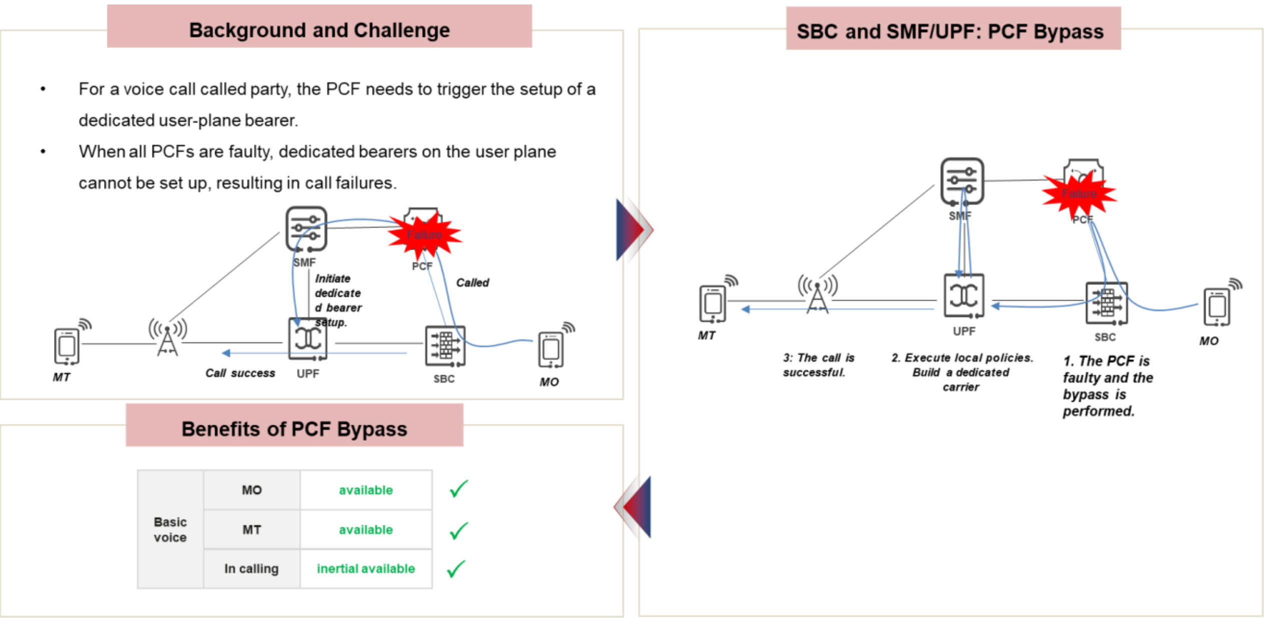
Note1: Newly registered subscribers are excluded.

Source: Thoth Advisory

Figure 36 depicts a scenario where a failure occurs in the PCF (Policy Control Function) which provides policy rules for control plane functions. This includes network slicing, roaming and mobility management. 4G LTE also has a PCF but not with network slicing. Under normal operation, when a voice call is placed, the PCF needs to trigger the setup of a dedicated user-plane bearer. When all PCFs are faulty, dedicated bearers on the user plane cannot be set up, resulting in call failures. One potential solution is as follows:

1. The SMF (Session Management Function) sends a local low-priority PCC policy to the UPF in advance.
2. When detecting that all PCFs are faulty, the SBC (Session Border Controller) performs voice service survival and sends RTCP voice packets to the UPF. The SBC is a network element deployed to protect SIP based voice over Internet Protocol networks. Early deployments of SBCs were focused on the borders between two service provider networks in a peering environment.
3. The UPF detects the SBC media-plane packet through DPI (Deep Packet Inspection and triggers the voice dedicated bearer setup procedure based on the packet information and local configuration.
4. The SMF rolls back the subscriber to a local PCC subscriber and activates the dedicated bearer by using the local low-priority PCC policy.

Figure 34: PCF Bypass to ensure VoNR/VoLTE calling availability



Background and Challenge

- For a voice call called party, the PCF needs to trigger the setup of a dedicated user-plane bearer.
- When all PCFs are faulty, dedicated bearers on the user plane cannot be set up, resulting in call failures.

SMF
Failure
PCF
Initiate dedicate d bearer setup.
Called
MT
Call success
UPF
SBC
MO

Benefits of PCF Bypass

Basic voice	MO	available	✓
	MT	available	✓
	In calling	inertial available	✓

SBC and SMF/UPF: PCF Bypass

SMF
Failure
PCF
MT
UPF
SBC
MO
3: The call is successful.
2. Execute local policies. Build a dedicated carrier
1. The PCF is faulty and the bypass is performed.

Source: Thoth Advisory

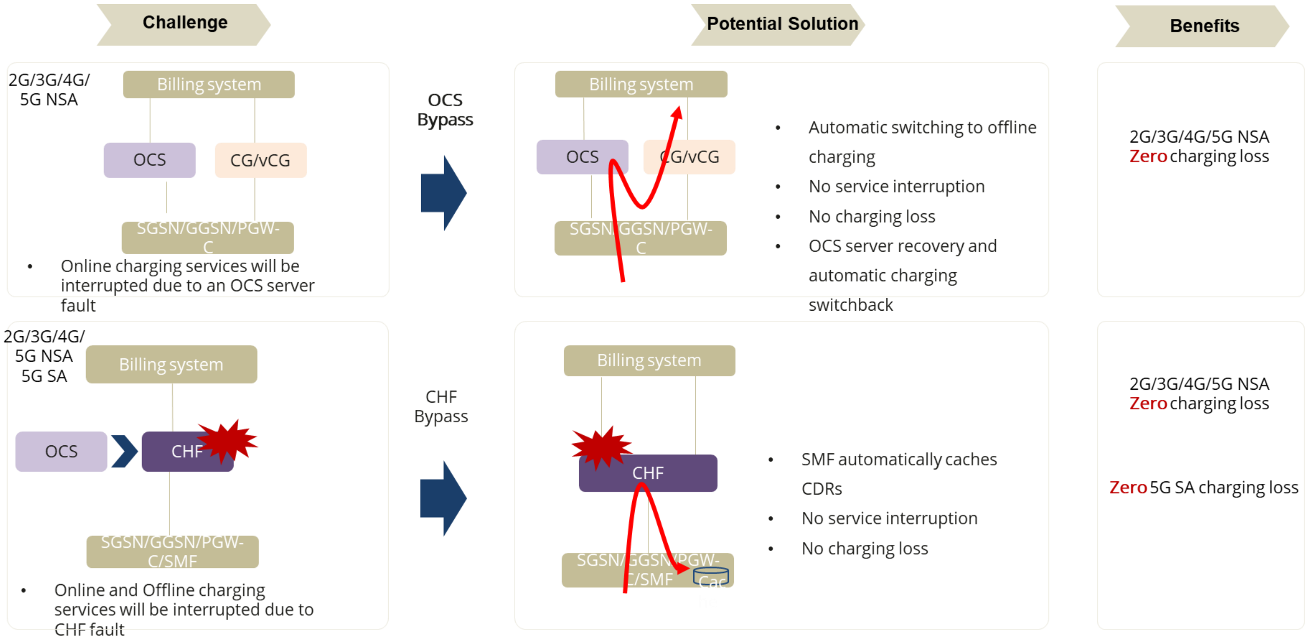
Figure 37 depicts a scenario where the Online Charging System (OCS) experiences a link failure. If a link to an OCS becomes faulty, the charging for subscribers is switched to offline.

1. A PGW-C and OCS link fault causes the charging type of an online charging subscriber to change to offline charging.

2. When the link recovers, the PGW-C immediately sends CCR-I messages to the OCS to establish a session for online charging restoration.

3. If another reason causes the charging type of an online charging subscriber to change to offline charging and later the link recovers, the PGW-C regularly sends CCR-I messages to the OCS to establish sessions for online charging restoration.
4. Other faulty scenarios include: The link between the PGW-C and OCS is functioning properly, but the PGW-C does not receive any CCA message from the OCS in a specified period of time after sending a CCR message. The link between the PGW-C and OCS server is functioning properly, and the PGW-C receives a CCA message that contains a result code indicating an OCS overload or failure.

Figure 35: OC/CHF bypass ensures continuity



- Online charging services will be interrupted due to an OCS server fault

- Automatic switching to offline charging
- No service interruption
- No charging loss
- OCS server recovery and automatic charging switchback

2G/3G/4G/5G NSA
Zero charging loss

- Online and Offline charging services will be interrupted due to CHF fault

- SMF automatically caches CDRs
- No service interruption
- No charging loss

2G/3G/4G/5G NSA
Zero charging loss

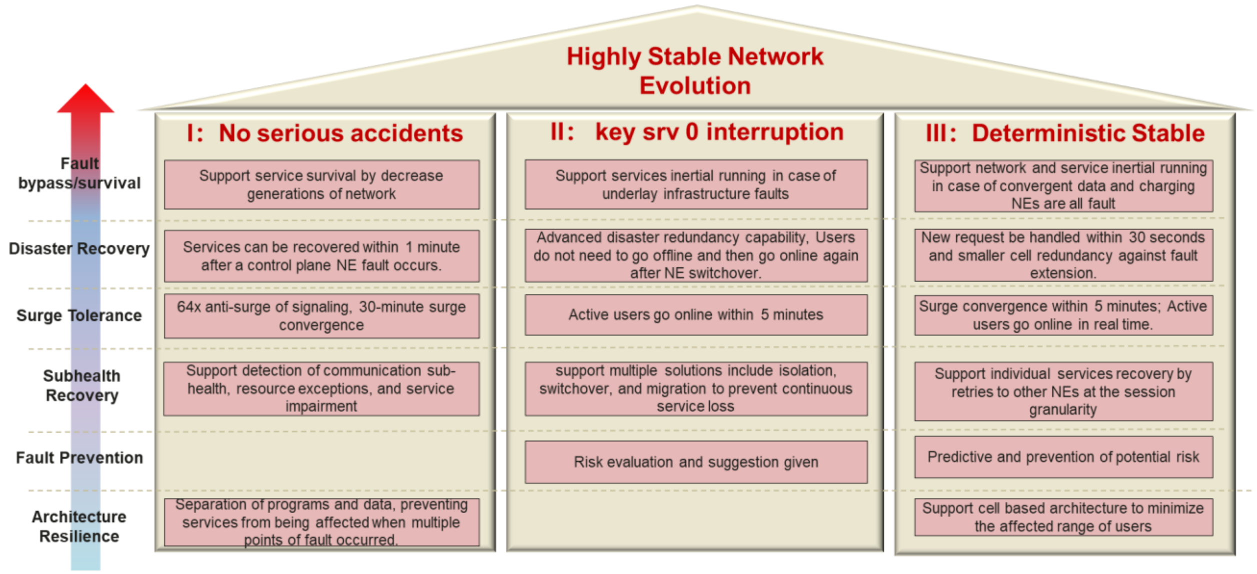
Zero 5G SA charging loss

Source: Thoth Advisory

Developing a Framework for 5G Core Reliability and High Stability

Figure 38 provides a framework comprising actions and targets for mapping the six (6) dimensions against three network states: (1) no serious accidents, (2) Key service interruption and (3) deterministic stable. Figure 39 illustrates the concept of Service Impairment Awareness method which is based on the fundamental premise that NEs be designed with service impairment awareness, which means that NEs need to expose their telemetric data so that KPIs can be tabulated in real-time or near-real time.

Figure 36: Framing High Reliability in terms of 6 dimensions


Highly Stable Network Evolution
I: No serious accidents
II: key srv 0 interruption
III: Deterministic Stable
Fault bypass/survival
Disaster Recovery
Surge Tolerance
Subhealth Recovery
Fault Prevention
Architecture Resilience
Support service survival by decrease generations of network
Support services inertial running in case of underlay infrastructure faults
Support network and service inertial running in case of convergent data and charging NEs are all fault
Services can be recovered within 1 minute after a control plane NE fault occurs.
Advanced disaster redundancy capability, Users do not need to go offline and then go online again after NE switchover.
New request be handled within 30 seconds and smaller cell redundancy against fault extension.
64x anti-surge of signaling, 30-minute surge convergence
Active users go online within 5 minutes
Surge convergence within 5 minutes; Active users go online in real time.
Support detection of communication sub-health, resource exceptions, and service impairment
support multiple solutions include isolation, switchover, and migration to prevent continuous service loss
Support individual services recovery by retries to other NEs at the session granularity
Risk evaluation and suggestion given
Predictive and prevention of potential risk
Separation of programs and data, preventing services from being affected when multiple points of fault occurred.
Support cell based architecture to minimize the affected range of users

Source: Thoth Advisory

Figure 37: Service Impairment Awareness Method

User access	✓5G initial registration: AMF/UDM/PCF/SCP ✓Default data bearer setup: AMF/SMF/PCF/UPF ✓Default voice bearer setup: AMF/SMF/PCF/UPF ✓IMS registration: UPF/SBC/CSCF/UDM	✓ Success rate: 5G initial registration, default data bearer setup, default voice bearer setup, and IMS registration ✓ Number of failures: 5G initial registration, default data bearer setup, default voice bearer setup, and IMS registration

Data service	✓ Connected data transmission (uplink and downlink): UPF ✓ Migration from idle to connected (uplink): AMF/SMF/UPF ✓ Migration from idle to connected (downlink): UPF/SMF/AMF ✓ Bearer session establishment: AMF/SMF/PCF/UPF	✓ Success Rate: AMF Service Request Success Rate/SMF Session Setup Success Rate/UPF Session Activation Success Rate/PCF Session Authorization Request Success Rate, UPF Downlink Forwarding Success Rate/SMF Data Notification Receive Success Rate/AMF Paging Request Success Rate ✓ Number of failed service requests: Number of rejected AMF service requests/Number of failed SMF session establishment/Number of failed UPF activations, Number of failed UPF downlink forwarding requests/Number of failed AMF paging requests/Number of failed SMF data notifications received
Voice call	✓ Dedicated bearer setup for voice services: AMF/SMF/PCF/UPF ✓ Voice calling: UPF/SBC/CSCF/ATS ✓ Voice callee: UDM/ATS/CSCF/SBC/UPF	✓ Success rate: voice dedicated bearer setup/PCF session authorization/voice calling/voice called ✓ Number of failures: voice dedicated bearer setup, PCF session authorization, voice calling, and voice called
Switchover	✓ Voice handover: AMF/SMF/UPF ✓ Data handover: AMF/SMF/UPF	✓ Success rate: Inter-eNodeB handover success rate/Inter-AMF handover success rate/5G-to-4G handover success rate/4G-to-5G handover success rate ✓ Number of failures: Number of inter-eNodeB handover failures/Number of inter-AMF handover failures/Number of 5G-to-4G handover failures/Number of 4G-to-5G handover failures
IMS SMS	✓ SMS origination: UPF/CSCF/IP-SM-GW/SMSC ✓ Short message termination: SMSC/UDM/IP-SM-GW/ CSCF/UPF	✓ Success Rate: SMS Origination/SMS Termination ✓ Number of failures: SMS origination/SMS termination
New call	✓ Calling party in a new call: PS/IMS/NCP/UMF ✓ New callee: IMS/PS/NCP/UMF ✓ Audio and video anchoring: third-party service AS/NCP/ATS/UMF/media capability platform ✓ Downloading the applet list: ATS/NCP/ATS/UMF (establishing a DC channel) and SBC/UMF/NCP (The UE obtains the applet buffered during startup from the NCP through the UMF based on the DC channel.)	✓ Success rate: new call calling, new call called, voice and video anchoring, and applet list download ✓ Number of failed calls: calling number of new calls, called number of new calls, voice and video anchoring, and applet list downloading

Best Practices in China

The sheer size, scale and geographic coverage (31 provinces) of the subscriber market in China has necessitated the need for an effective reliability strategy that is quantifiable and can be continually improved. The strategies can be summarized as follows and depend on two basic topologies: (1) active-active and (2) redundant connections for the NRFs:

- Active/standby + single layer + full interconnection between eight large areas
- Active-active+two-layer+31 provincial and municipal L-NRFs interconnected with six pairs of H-NRFs
- Active-active+two-layer+eight-region L-NRF aggregation to a pair of H-NRFs in China
- Inter-large DR Inter-province disaster recovery within a large area: Shorten bearer network paths and reduce fault points. Reduce the impact area and affected provinces by DC faults by half. Reduce the signalling connection delay on the RAN side.
- Creation of KPI baselines such as (excluding user causes)
    - The success rate of initial registration on the ToC AMF
    - ToC SMF PDU session establishment success rate
    - ToB AMF initial registration success rate
    - ToB SMF RADIUS Authentication Success Rate
    - VoNR voice network call completion rate

Another area where the three operators in China have been proactive with regards to network reliability is in the fault handling evaluation systems which comprise fault determination accuracy, handling measure effectiveness, and troubleshooting duration. Ideally in the future, there will be more and more automation involved in the root cause analysis but reality on the ground is that maintenance personnel still play a critical role. Thus, the operators in China have learned that they also a need to monitor maintenance personnel's effectiveness and their ability to locate fault causes.

Appendix

Classical Availability Theory: MTBF and MTTR

Modern mathematical theory of reliability dates back to the late 1960s. Reliability analysis has become an important branch of applied mathematics since then. Despite huge advances in reliability research such as multi-state reliability modelling, multivariate reliability function, and the introduction of survival analysis and competing risks, the basic definition of reliability has remained in intact.

Mathematically, reliability is expressed as a probability using the following framework: We are concerned with a device that fails at an unforeseen or unpredictable random time (or age) T>0, with distribution function F(t)

$$F(t) = \mathscr{P}\ (T>1)$$

where $\mathscr{P}$ is the probability, T is the lifetime or failure time, and the probability distribution function (pdf) is f(t). The reliability $\mathscr{R}$(t) is then defined as R (t) = 1 – F(t). The failure rate λ(t) is defined as the ratio of pdf f(t) to reliability:

$$\lambda(t) = f(t)/\ \mathscr{R}(t)$$

The failure rate λ(t) is sometimes referred to as the *instantaneous failure rate*. The cumulative hazard rate function H(t) is related to the failure rate and the Reliability $\mathscr{R}$ as follows:

$$H(t) = \int_0^t \lambda(s)ds = -\,log_e[R(t)]$$

Thus, the Reliability R(t) is the exponential of the Hazard function H(t).

<u>Mean Time Before Failure (MTBF)</u> is a measure of but not guarantee of the reliability of a system or component. MTBF is an important factor in developing maintenance strategies. MTBF is usually calculated as an average value that can be used to estimate the expected service life of a system or component. MTBF is expressed as (See the Appendix)

$$\text{MTBF} = \textstyle\sum\{\text{ Start of downtime -Start of Uptime}\}/\text{Number of failures} = \int_0^\infty R(t)\,dt$$

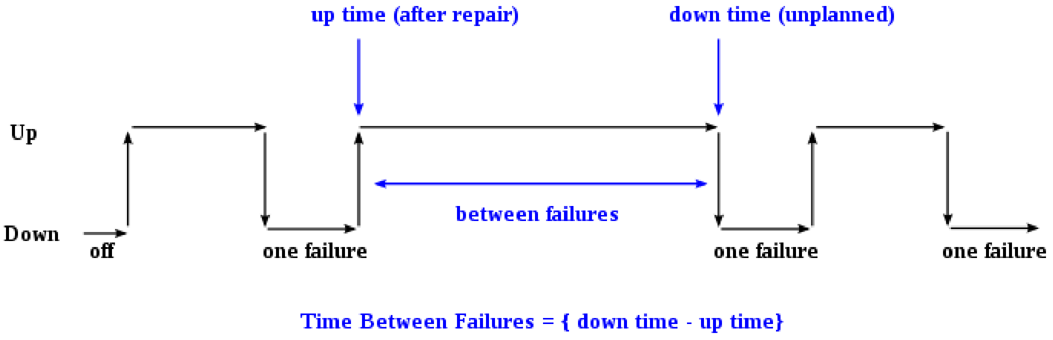
where $\mathscr{R}$ (t) is the reliability function. MTBF can also be written as the inverse of the failure rate λ(t):

$$\text{MTBF} = 1\ /\ \lambda\ (t)$$

Once the MTBF of a system is known, the probability P, assuming a constant failure rate, that any particular system will be operational for a given duration can be inferred from the reliability function of the exponential distribution RT(t) = $e^{-\lambda t}$. The probability that a particular system will survive to its MTBF is 1/ *e, or* ~*37%.* If a system for example has two components and if the failure of one causes the network to fail then this topology is considered to be serial. If both components must fail to cause the system to fail then the network is said to be in parallel. Assuming the MTBF of each component c_1 and c_2 is known then the total MTBF is similar to a parallel RC electrical circuit and is given by:

$$\text{MTBF (c1, c2)} = \frac{1}{\frac{1}{m(C_1)}+\frac{1}{m(C_2)}} = \frac{m(C_1)m(c_2)}{m(C_1)+m(C_2)}$$

Figure 39:Time Between Failures = {Down time - Up Time}



Note: The down time is the instantaneous time the system went down and is after (i.e. greater) then the moment the system went up.

Another important network concept is the Mean Time to Repair (MTTR) which is the average time it takes to repair or recover a system (this includes testing time). An example is the time it takes for submarine cable systems to be repaired when an earthquake or fishing boat net or anchor breaks the optical fiber cable. Subsea repairs can take 3-5 days in a good case and several weeks a worst case to carry out once the ship is on site and thus can be highly disruptive to the global internet. Due to the high cost of construction subsea cable systems (tens of millions of $ for branch links) alternative paths are not always economical.

Contact Us

If you have any more questions regarding our research, please contact us:

Principal
Ray Er
ray.er@thothadvisory.co

Disclaimer: All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher, Thoth Advisory. The facts of this report are believed to be correct at the time of publication but cannot be guaranteed. Please note that the findings, conclusions and recommendations that Thoth Advisory delivers will be based on information gathered in good faith from both primary and secondary sources, whose accuracy we are not always in a position to guarantee. As such, Thoth Advisory can accept no liability whatsoever for actions taken based on any information that may subsequently prove to be incorrect.