ASGC Incident Report

Simon C. Lin

Eric Yen

Academia Sinica Grid Computing

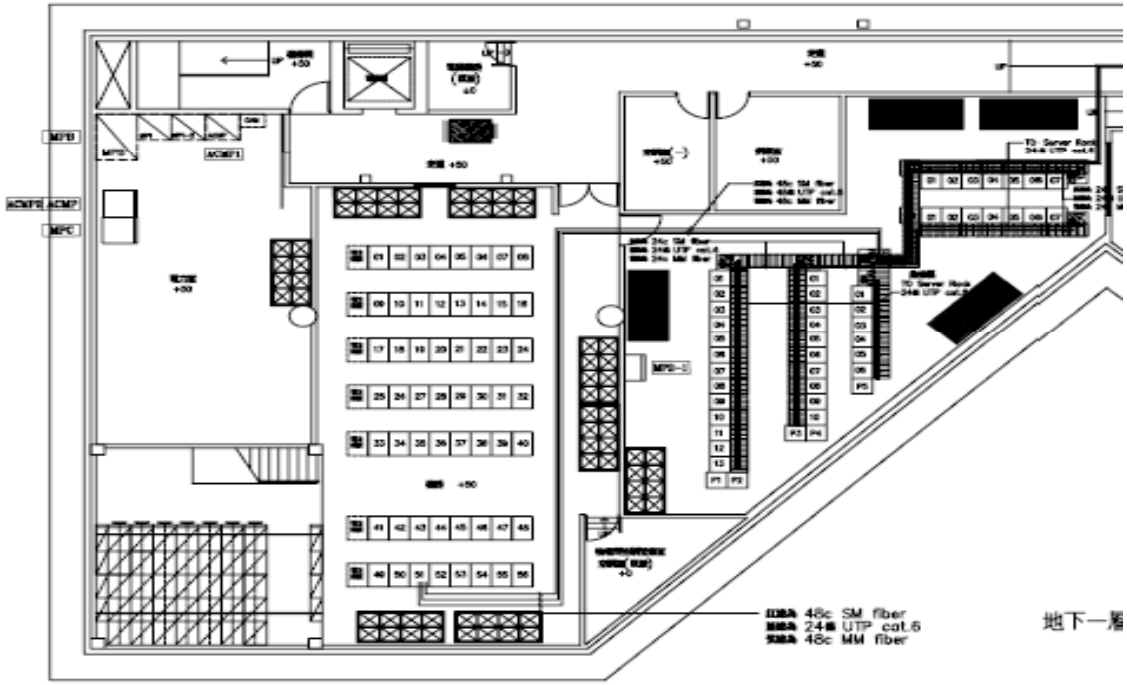
11 March 2009, GDB, CERN

ASGC Data Centre

Background Information

Capacity & Expansion

- Total Capacity
  - 2MW, 330 tons
  - 100 racks
- 2007 Expansion
  - 43 Server Racks
  - Cooling: 120 Tons: N+1
  - 13 kVA per rack
  - Improved efficiency
    - Overhead cable trays
    - Hot Cold Aisle
    - Minimize cool air leakage areas
- 2008 Expansion
  - 2MW generator
  - UPS expansion
  - Cooling system for C#3 (tape system)



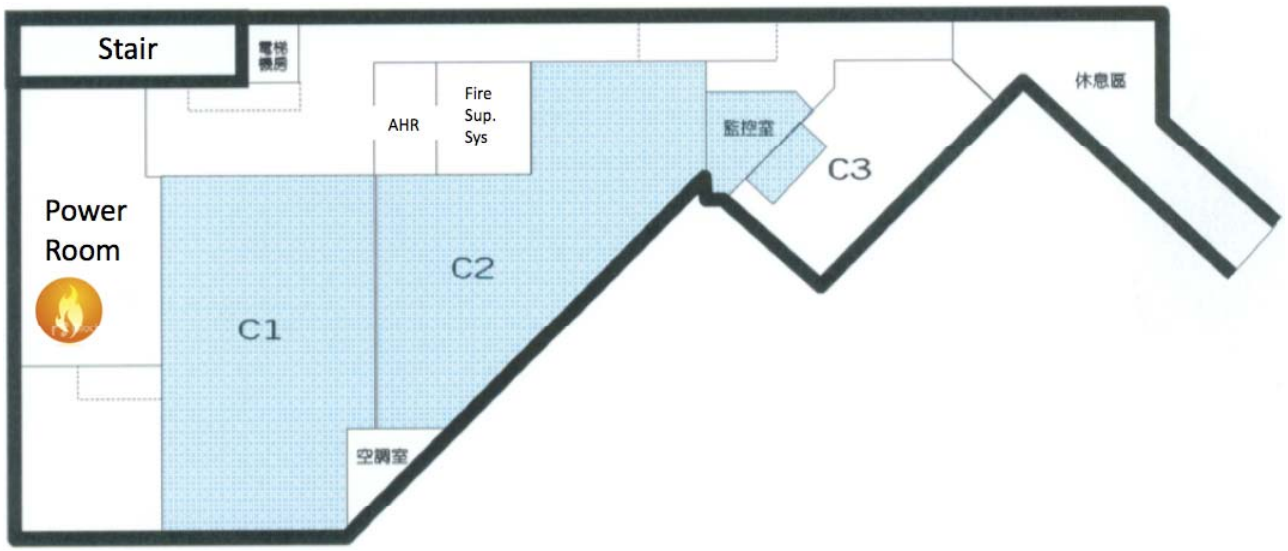
Power Room – UPS & Panel (I)

- Two 400KVA online UPS
  - FRAKO Aktives Filters OSF
  - SPB Panel
- Battery (RHS) series
  - 150Ah Leadline

Power Room – UPS & Panel (II)

- Four 200KVA online UPS
  - IGBT inverters
  - N+1 redundancy
- Battery (RHS) series
  - NPA-120-12 Yuasa

Data Centre Incident @ ASGC



Event Scene Investigation

- 16:53 on 25 February (Taipei Time), fire lasted less than 3 hours:
  - SMS sent out “UPS power lost” and power generator did not start
  - Smell and visible of smoke, and then fire at 2nd column of battery rack. Electric arc also viewed in the middle battery racks.
  - Fire spot (V shape) on the far-end wall, and near-end battery rack was also burned out → fire spread to left and right within the middle battery racks, and up to the ceiling
  - But no fire burning evidence at racks in the other two ends

Burnt-out UPS Battery

Network Recovery Plan & Action

- Within the incident investigation at day 1 & 2, we looked for the possible domestic backhaul routes (due to the different submarine service providers) and machine rooms to install our core router (Juniper M320)
- Permitted to move the router at day 3, we started to clean the facilities and inspected the router functions in 12 hours. Fortunately, all of the major parts and interface modules are working well. However, we replace the cabinet and power supply parts.
- We restored all international links: 10Gbps TPE-SARA, 2.5Gbps TPE-CHI-SARA, 2.5Gbps TPE-JP, 2.5Gbps TPE-HK and 622Mbps TPE-SG at 5:30 pm, and the domestic links at 7:30 pm in day 4 (28 February)
- We even provided the transit service via Singapore for the APAN Kaoshiung Meeting Demonstration at 2~6 March
- Can quickly relocate the network PoP when the machine room is ready

Grid Service Recovery

- Objectives: Minimize impacts to Tier1 and Tier2 services @ ASGC
- Before DC recovered
  - Resume Core Services First in another place
    - Networking, ASGCCA, GStat, VOMS, APROC, TPM, DNS, mail/list back online at 27 Feb.
    - BDII, LFC, FTS, VOMRS, UI, DPM recovered at 7 March
  - Resume Data Catalog and Access Service
  - Resume Degraded Tier1 and Tier2 Services
- Alternatives
  - Distributed Racks: Tier2 ready @ IES at 10 March
  - Co-locate Tier1 Resources at IDC: scheduled at 15 ~ March (?)

Recovery (I)

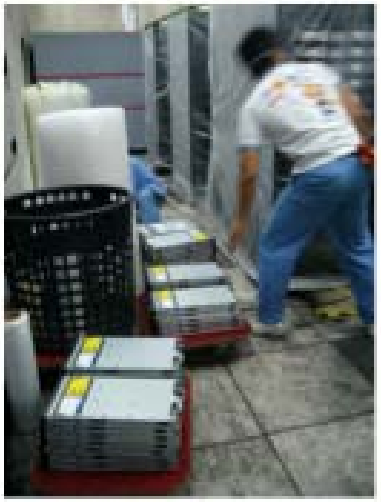
27 Feb: Power system inspection and temporary power installed

27, 28 Feb: DC Cleaning and packing

28 Feb: Facility investigation for recovery plan (AHU)

28 Feb – 2 March: Rack protection and Ceiling removal

Recovery (II)



2-5 March: Move out all facilities in DC

2-13 March: Clean up all ICT facilities & test

Protect Racks from Dust

Ceiling Removal

Dust Cleaning

Move out all facilities for cleaning

Dust cleaning in and out

Container as temporary dry storage

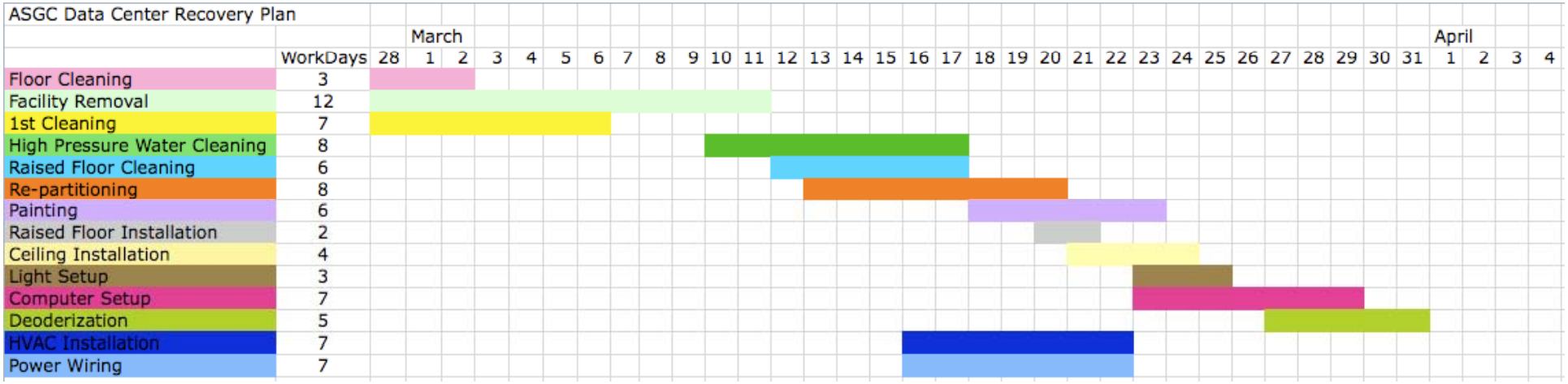
Impact Analysis (I)

- Air Quality and safety of DC
  - Carbon and soot particle density is 3 times of normal in the 1st week after accident
  - Dioxin? Examination is under way
- Power System Failure
  - UPS replacement
  - Busway replacement: 2 weeks for production and installation
  - Re-wiring
- HVAC system
  - Pipeline replacement, ventilation re-design, defect fix
- Fire Prevention System: system verification, N2 refill
- Elevator: defect fix, verification
- Monitoring System: re-design and installation
- DC Construction (Architecture): from ceiling to floor, re-partitioning
- Computing & Storage System: clean dust particle ins and outs

Impact Analysis (II)

- WLCG Services
  - Computing & Operation model robust enough to sustain from one or more T1s failure?
  - Data replication strategy
- Experiment Computing Model
  - Is the computing model flexible and resilient enough to recover from one or more T1s failure?
  - What's the impact to users? How to minimize it?
- Communication
  - Between site and WLCG, Experiments
  - WLCG and users

Recovery Plan



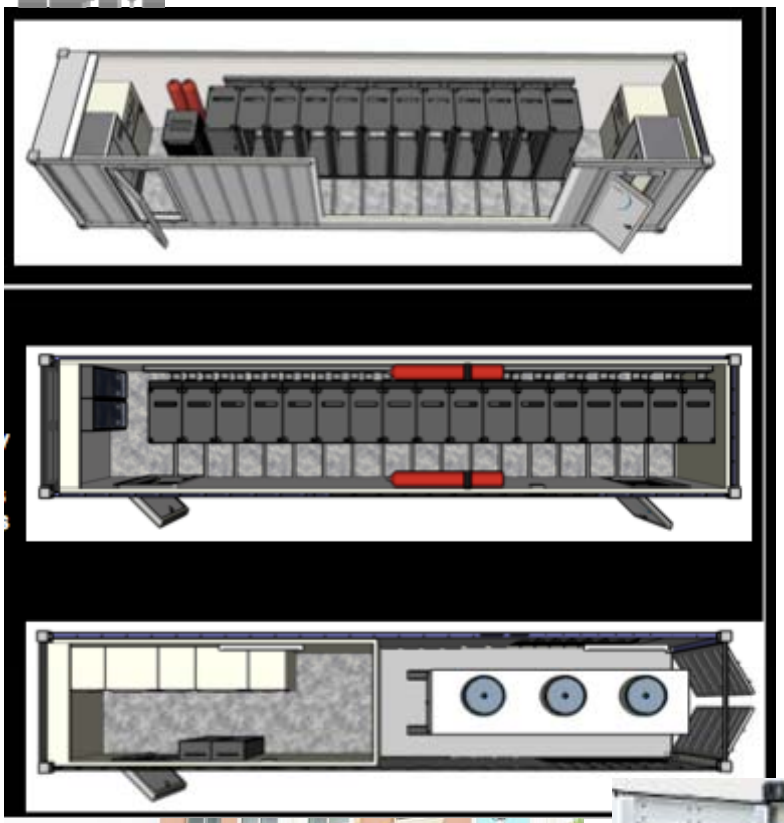
- DC Consultant will review the re-design on 11 March, schedule will be revised based on the inspection
- Tier1/Tier2 resources, alternatively, will be co-located at IDC for 3 months from 20 March hopefully (3 days for removal and installation)

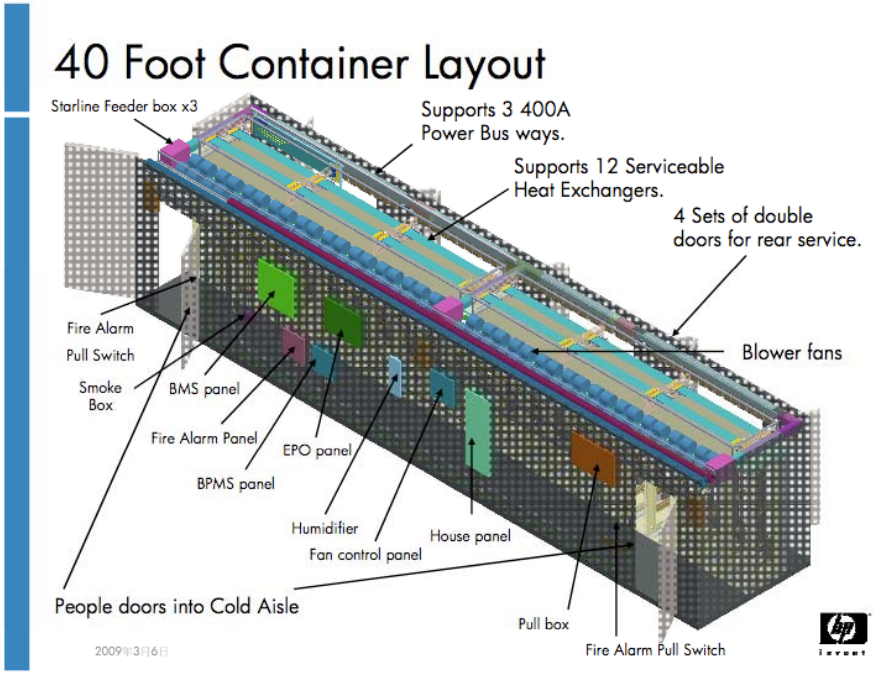
Lessons Learnt

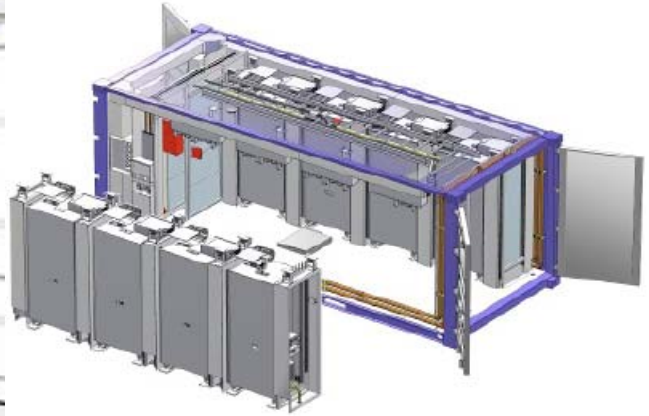
- DC Infrastructure Standards to comply with
  - ANSI TIA/EIA
  - ASHRAE thermal guideline for data processing environment
  - Guidelines for green data centers are available, e.g., LEED
- Which model UPS is most safe? Battery, Flywheel, Rotary or Combinations? Cost-effective is still an issue.
- UPS battery better to be located outside the DC (open space with enclosure) and keep the minimum capacity only.
- Disaster Recovery plan is necessary, for various risk levels for example.
- Routine Fire drill is indispensable.

Performance, Mobility and Customization DC Alternatives

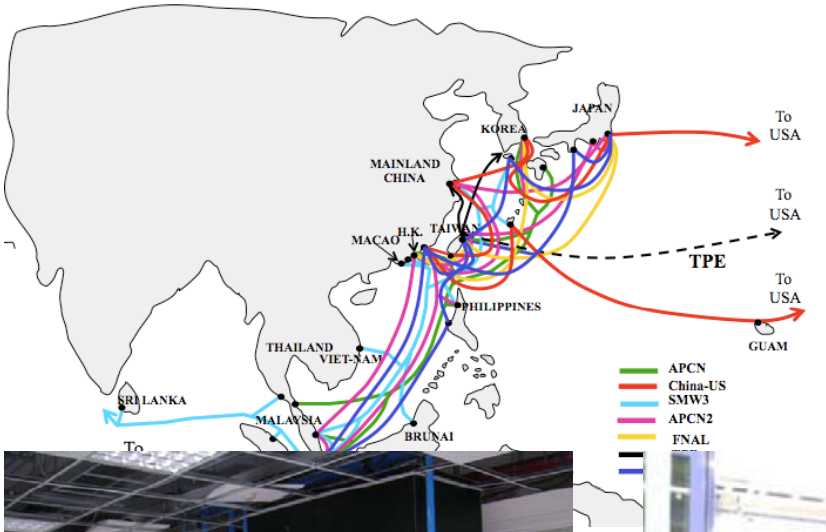
IBM







IDC Co-location



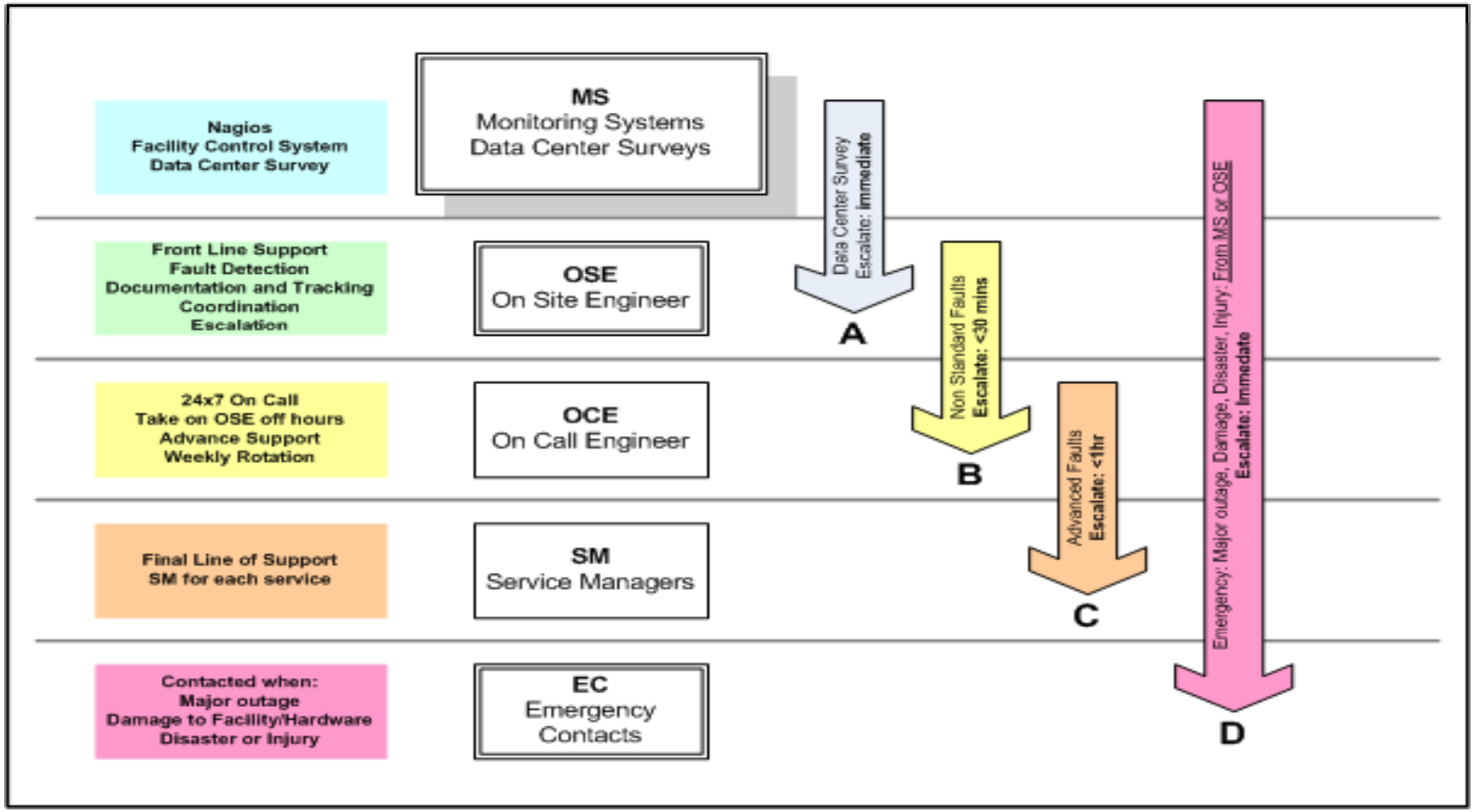
24x7 Services Implementation

- Will implement 24x7 on-site service 6-month ahead in April.
- On Site Engineers (OSE): 24 hours a day, 7 days a week
  - On-site inspection of data center
  - Problem management:
    - Detect faults and issues with aid of monitoring system
    - First line troubleshooting
    - Problem coordination of open issues, and escalation
    - Follow ups
  - Ticket management
- On Call Engineers (OCE): 24x7
  - Support OSE to resolve issues
  - Provide coverage when OSE not available
- Service Manager (SM): 24x7
  - Service experts and last line of support to resolve complex issues
- Emergency Contacts (EC): 24x7
  - a major outage occurs
  - damage occurs to hardware and Data center facilities
  - major physical damage, disaster or injury

Service Classes

- Foundation services: 99.9%
  - Data center, DB, Network, DNS
- Critical services: 99%
  - T1, T2, CA, EGEE, Ticketing System, Nagios, Mail
- Best effort: Next business day
  - UI, RGMA, IT, etc.

Escalation


Nagios
Facility Control System
Data Center Survey
MS
Monitoring Systems
Data Center Surveys
Front Line Support
Fault Detection
Documentation and Tracking
Coordination
Escalation
OSE
On Site Engineer
24x7 On Call
Take on OSE off hours
Advance Support
Weekly Rotation
OCE
On Call Engineer
Final Line of Support
SM for each service
SM
Service Managers
Contacted when:
Major outage
Damage to Facility/Hardware
Disaster or Injury
EC
Emergency
Contacts
Data Center Survey
Escalate: immediate
A
Non Standard Faults
Escalate: <30 mins
B
Advanced Faults
Escalate: <1hr
C
Emergency: Major outage, Damage, Disaster, Injury: From MS or OSE
Escalate: Immediate
D

Incident Summary

- Damage Analysis: fire was limited to the power room only
  - Severe damage of UPS
  - wiring of power system, AHR
  - Smoke dust pervaded and smudged almost everywhere, including computing & storage systems
- History and Planning
  - 16:53 on 25 Feb. UPS battery burning
  - 19:50 on 25 Feb. Fire extinguished by Fire department
  - 10:00 on 26 Feb. Fire scene investigation by Fire department
  - 15:00 26 Feb. to 23 March, DC cleaning, re-partitioning, re-wiring, deodorization, and re-installation
    - from ceiling to ground under raised floor, from power room to machine room, from power system, air conditioning, fire prevention system to computing system
    - All facilities moved outside to cleaning
  - 23 March, Computing System installation (originally planned time without co-location)
  - 23 March ~ 9 April, Recovery of Monitoring, Environment control and Access control system
  - Alternatively, co-locate Tier1/Tier2 resources at IDC for 3 months from 20 March hopefully