NARLabs

High Performance Storage and Execution Architecture

SEVENTH FRAMEWORK PROGRAMME

fish 4 knowledge

國家實驗研究院 National Applied Research Laboratories

NARL

NCHC

Objectives

O1: Achieve scalable long term real time capturing and buffering for multiple undersea video stream.

(Data)

O2: Build a Tera-scale data service platform consisting of repositories for the video data, for the metadata, for the processed data and for the live stream data, and a computational cluster to support analysis.

(Compute &Store)

O3: Achieve high performance data store and computational access for the data service platform.

(Query Performance)

NARLabs

O1: DATA

A – 1 : Video Summary

Video Data Collection Status in NCHC

Video Format	Video Bitrate	Site Name	# of Video in Storage	# of Video Record in DB
FLV	200K/480K/ 1M/2M	All Sites	685,607	662,804
MPEG4	5M	NPP-3	41,977	none

10min per video record

Last Updated on 11/01/2013

NARLabs

Video Data Collection Status in NCHC

Video Format	Video Bitrate	Resolution	Site Name	# of Video Record in DB
FLV	200K/480K/ 1M/2M	320x240	All Sites	200,004
FLV	200K/480K/ 1M/2M	640x480	All Sites	368,547

Resolution	<5 fps	5 ~ 8 fps	9 ~ 23 fps	24 fps	>24 fps
320x240	5,520	189,101	5,383		
640x480		90,653	12,356	264,421	1,117

Resolution	2 fps	4 fps	5 fps	8 fps	9 fps	10 fps	15 fps	20 fps	24 fps	25 fps	30 fps
320x240	337	5183	128903	60198	4178	1205					
640x480			41638	49015			4833	7523	264421	718	399

Last Updated on 11/01/2013
Recording video with 24 fps since 2011/04/27 06:00:00

Current Data Storage Resources Status in NCHC

	Size	~ 2013-07-16	~ 2013-10-27	
NAS_1	14 TB	8.2TB (60%)	8.3TB (61%)	Historical video storage
NAS_2	14 TB	9.0TB (66%)	11.0TB (76%)	
NAS_3	8.2 TB	695MB (1%)	1.2GB (1%)	
NAS_4	8.2 TB	2.1TB (26%)	2.1TB (26%)	
NAS_5	8.2 TB	7.8TB (96%)	7.8TB (96%)	
NAS_6	8.2 TB	3.6TB (44%)	3.6TB (44%)	
NAS_7	13 TB	-	-	VM NFS shared storage (damaged and moved to NAS_9 already)
NAS_9	107 TB	47TB (44%)	58TB (55%)	F4K data storage & video backup
Total Storage	180.8 TB	77.7 TB (43%)	90.8 TB (50%)	

Last Updated on 10/27/2013

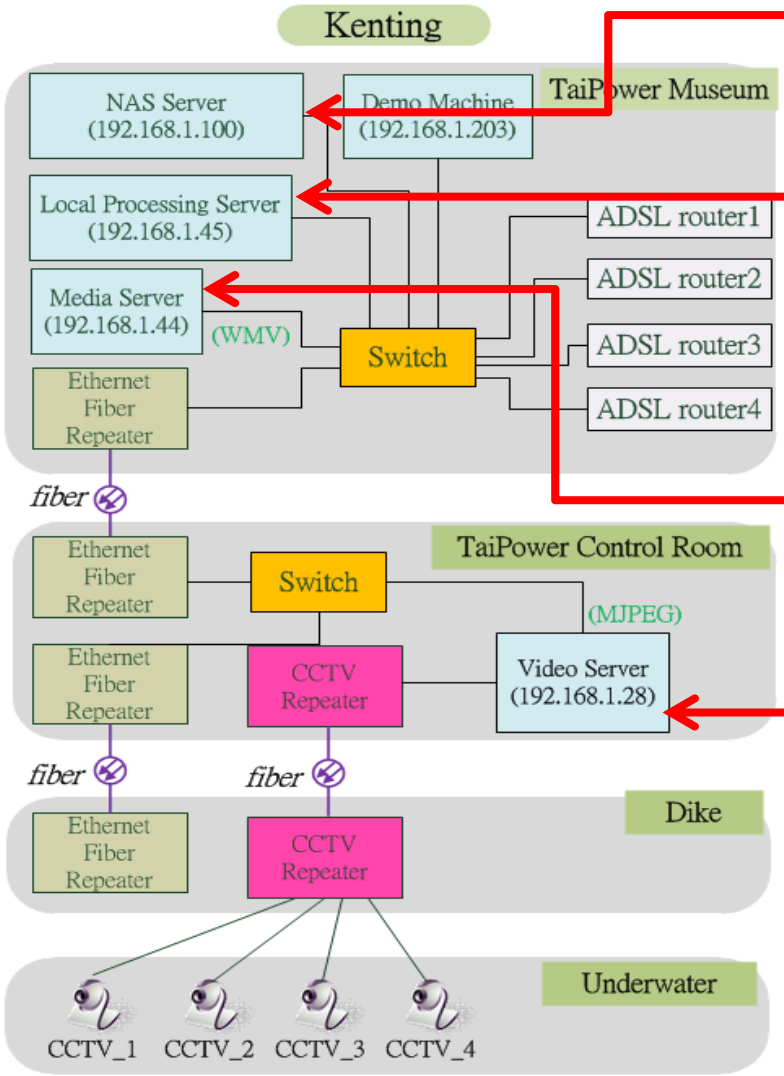
NARLabs

Native Motion JPEG Video for reference

- Provide about 10 Hours Native Motion JPEG (MJPEG) video from experiment on 5th of February. (around 9:30 AM ~ 19:00 PM)
- Each MJPEG video size is nearly 40 GB.
- The experiment was running at 10AM to 11AM and 2PM to 3PM.
- Split the MJPEG video to 5 ~ 10 minutes short videos. Each video size is about 250MB ~ 600MB
- Download Link:

http://gad249.nchc.org.tw/tom/tdw/ecodata/Site_A/Video/10_min/

Multi-stage Data Streaming for the Monitoring System



5. Local NAS Storage
7.7TB RAID 5 Storage for storing 5M mpeg4 video

3. VLC
Capture 6M mpeg stream and convert into 6M mpeg4 video file (10 mins)

4. Ffmpeg
Convert 6M mpeg4 video to 5M mpeg4 video File.

2. VLC
Capture Motion JPEG stream and convert into mpeg4 format Stream (6M, deinterlaced).

1. Motion
Capture video signal from CCTVs and convert into Motion JPEG stream.

NARLabs

Ganglia Automatic Report System

- Our streaming system consists of many machines in Kenting, Taichung, and Hsinchu. Problems may happen to these machines, internet, power supply, etc. We need to handle these issues properly and rapidly for the **highly distributed system**.
- A **Ganglia automatic report system** was implemented to send emails to corresponding persons for notification of newly damaged, continuously broken, and repaired status.
- **Probing** every **5 minutes** to minimize the issues of recording failure.

O1: DATA

A – 2 : Corrupt Video Information

NARLabs

Video Download Error

- The failure of hard disks in NAS resulted in the inconsistence between database and physical file system.

Location	Counts	2009	2010	2011	2012	2013
cam-1 @ NPP3	288	88	1	129	70	
cam-2 @ NPP3	90			78	5	7
cam-3 @ NPP3	80	2	3	73	2	
cam-4 @ NPP3	85		3	79	2	1
cam-1 @ HoBiHu	4			4		
cam-2 @ HoBiHu	6			6		
cam-3 @ HoBiHu	18			18		
cam-1 @ LanYu	66	33	18	15		
cam-2 @ LanYu	20			20		

- Solution: sync both file system and DB regularly.

Synchronize Between DB and File System

- Backward SYNC
  - Read records in DB to check the existence of files
    - Not exists: delete the record in DB
    - Exists: modify size/timestamp in DB if needed
  - Modify 431 videos at 2013/7/13.
- Forward SYNC
  - Read physical video file in file system to check the record in DB exist or not.
  - Modify ~170K records at 2013/7/13

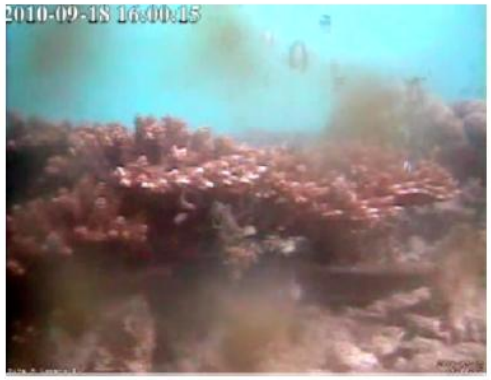
NARLabs

O1: DATA

A – 3 : Realtime Video Detection and Notification

Video Classification (~350K videos from 2009 to 2013)

Algae: 9.2%



Blurred: 33.5%

Complex Scenes: 4.3%

Encoding: 23.9

Highly Blurred: 13.9%

Normal: 12.9

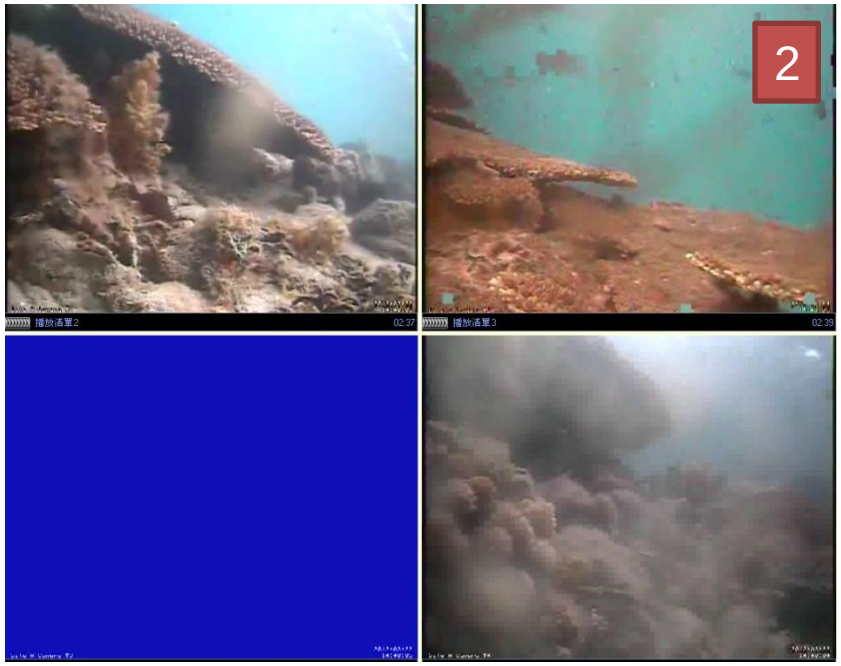
Unknown: 2.2%

Probably Causes of Abnormal Video

- Algae, Blurred, Highly Blurred
  - Some climatic factor like typhoon or heavy raining
  - Cleanness of camera lens
- System Errors
  - Blue screen: camera failure.
  - Gray screen: video capture card failure
  - Black screen: night
  - Mosaic/noise: internet traffic-jam, going to broken

Three Types of Video Stream

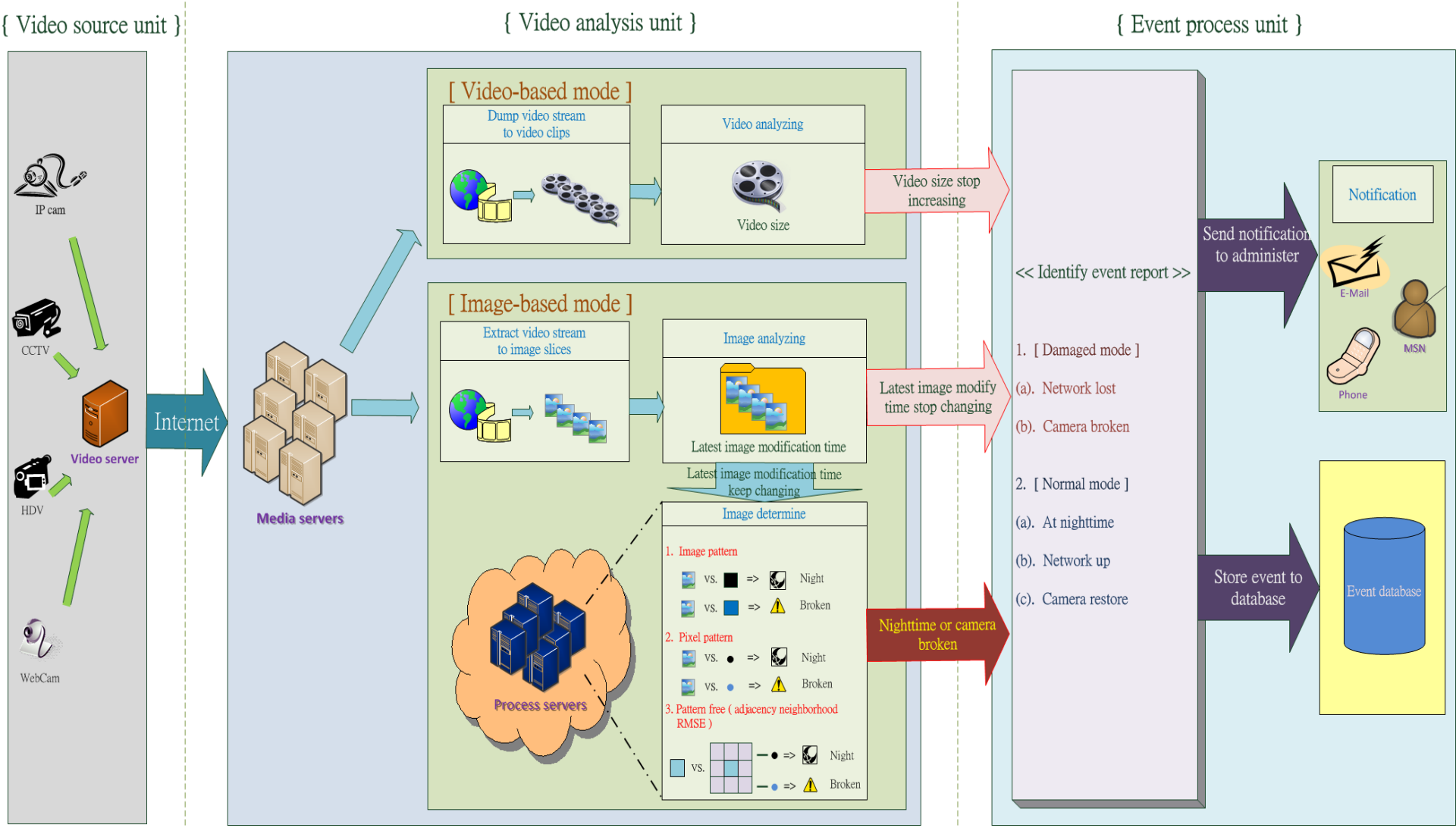
1. Normal
2. Broken in the left-bottom camera
3. Evening



Video-Slicing and Image-Analyzing Procedure

- Slicing video into images: *ffmpeg*
  - One frame per second.
  - To saving the computing time, sampling only 3 images (first/middle/last) for each video.
- Parsing each sliced image: *JPEGParser* program
  - JPEG_Filename
  - RGB_Boundary_Value: 000000 - FFFFFF
  - RGB_Boundary_Ratio: 0 - 100
  - RGB_Difference_Value: 0 - 255
  - Parser_Scheme: 1 – 3 (pixel-based, line-based, average)
  - RGB_Filename: save to a RGB ASCII

Detect & filter video from system errors (multistage data streaming)



NARLabs

O2: Compute & Store

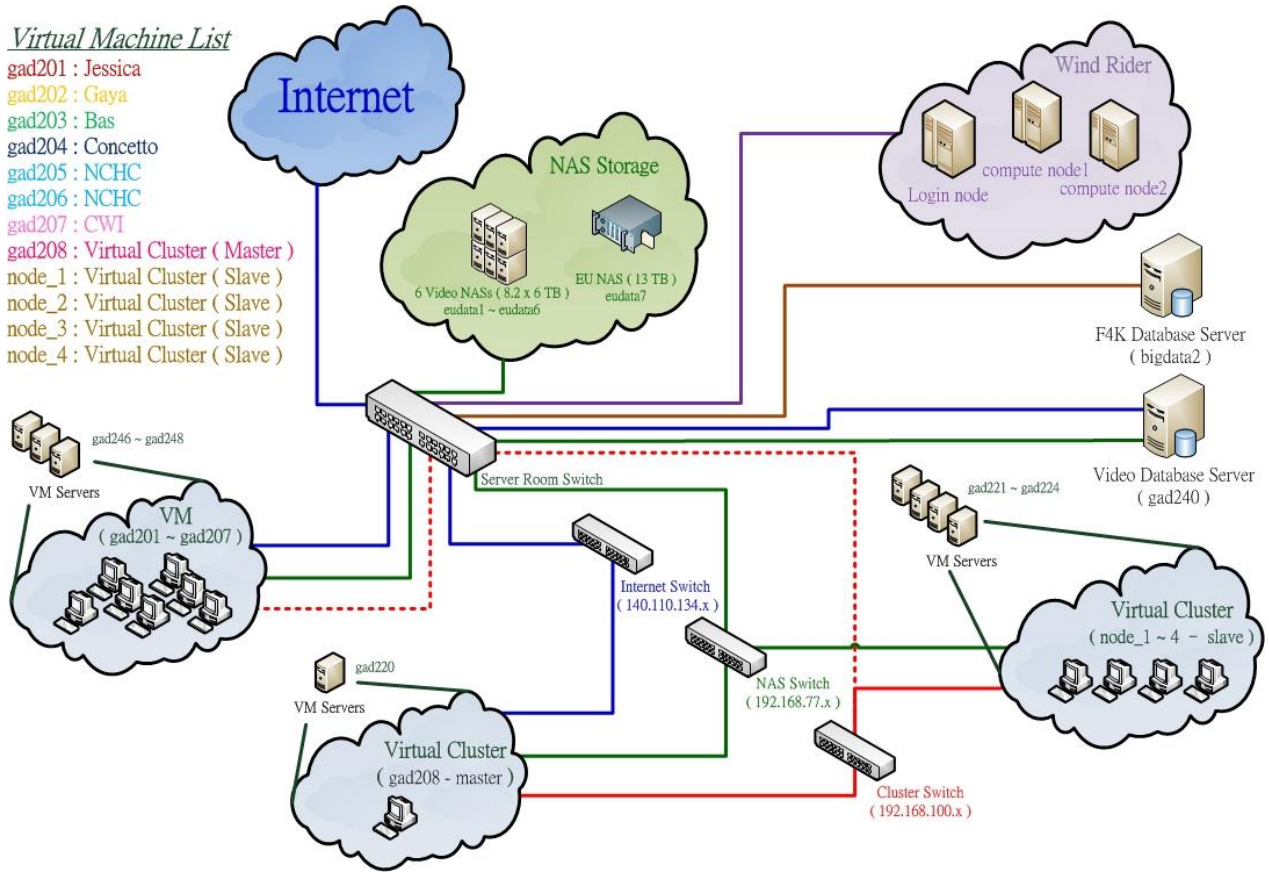
B – 1 : WindRider and VM Group

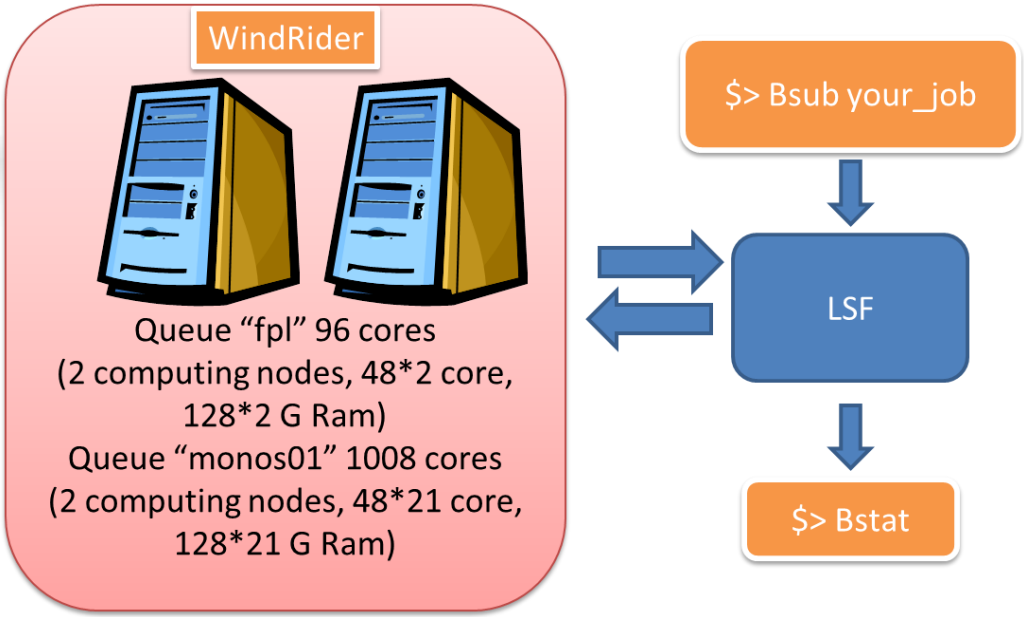
Computate Resources

- Computer systems
  - WindRider: 2 nodes, 96 CPU + additional on-demand 4.7 M core-hour
  - Experimental VM group.
    - Migration to Formosa 3 - Production Cloud.

- Summary of core-hours for 2013
  - 5.8M core-hour (WindRider) provided.
  - 3.5M core-hour practically consumed. (~400 core-year)
  - The support will be extended to 2015

Network/Resource Architecture for F4K

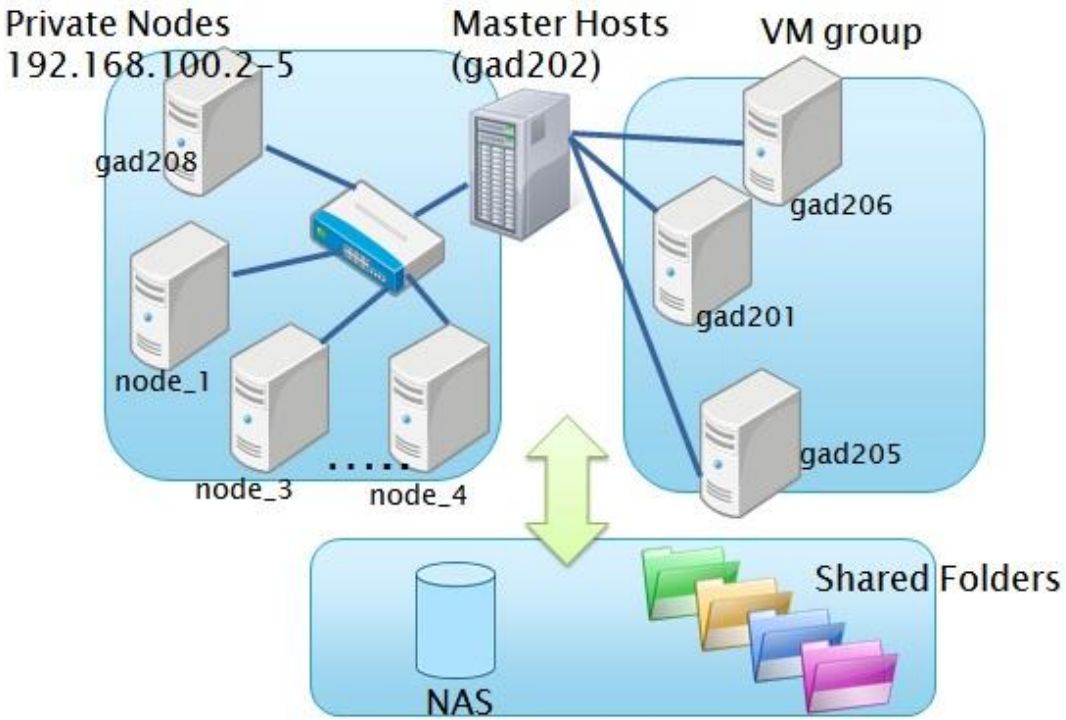




- The general purpose system uses the AMDR Opteron 6100 processors, and has a total of 8 compute clusters, 1 large memory cluster, and over 25,600 cores.
- It is a supercomputer that offers an aggregate performance of over 177 TFLOPS.

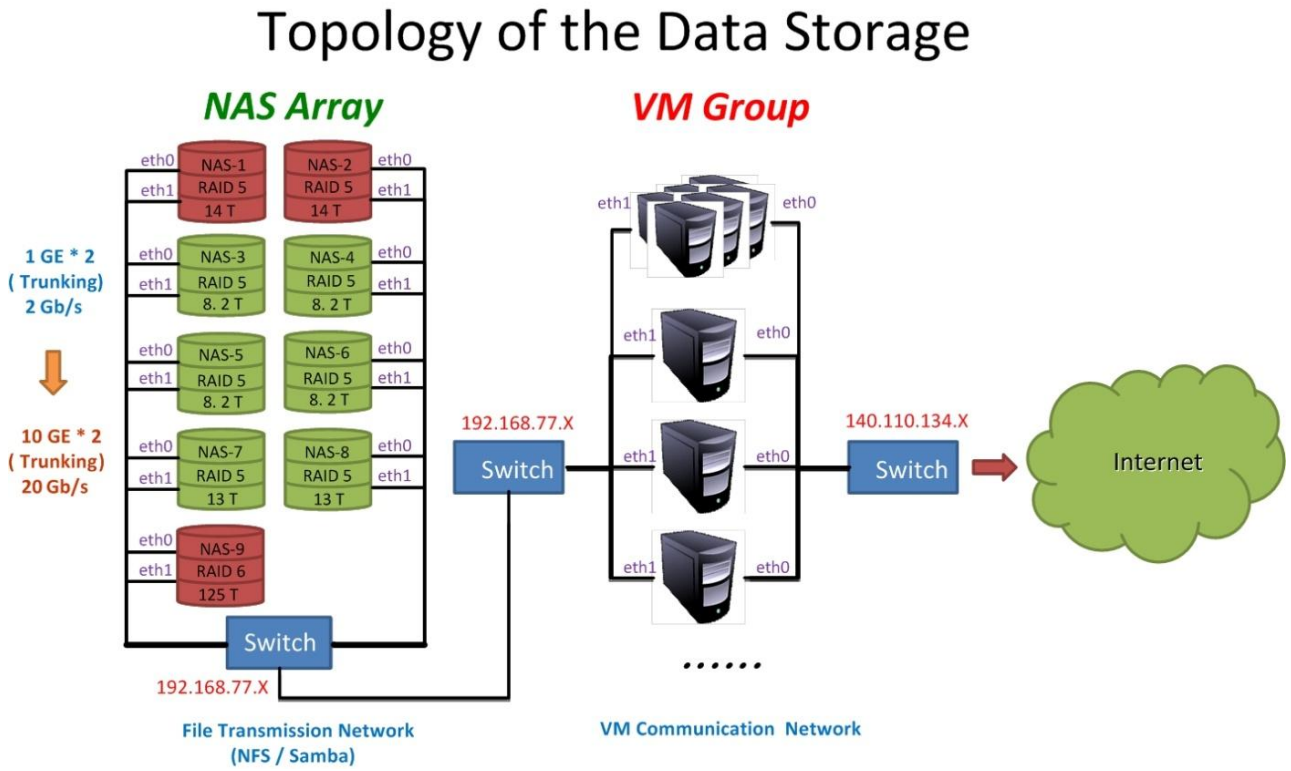
VM Group

- Architecture of VM Group



Data Storage

NAS Storage Topology - Private Storage Network



NARLabs

VM group backup to Formosa3

- VM group migrate to Formosa 3, as backup VM group for testing or additional computing resources.
- Current VM group state

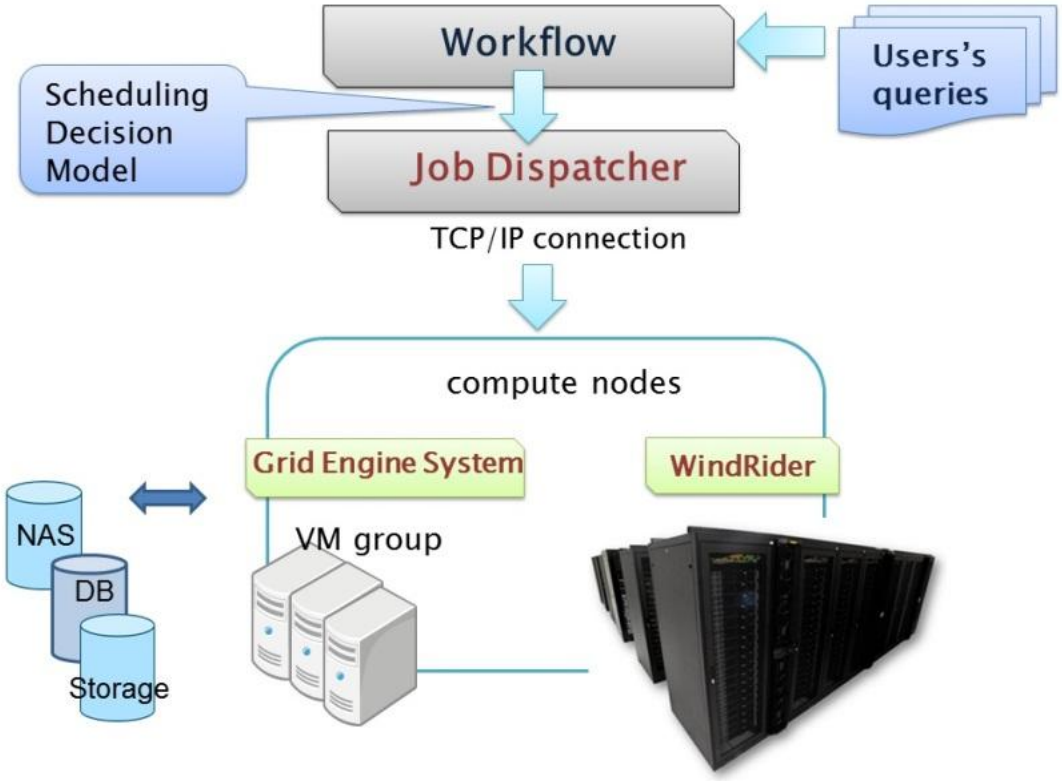
VM Name	Owner	CPU	Memory (GB)	Cluster Node	SGE Node
Gad201	Jessica	8	16		Slave
Gad202	Gaya	12	32		Master
Gad203	Bas	8	16		
Gad204	Concetto	8	16		
Gad205	NCHC	8	16		Slave
Gad206	NCHC	8	16		Slave
Gad207	CWI	8	16		
Gad208	NCHC	6	16	Master	Slave
Node_1	NCHC	6	6	Slave	Slave
Node_2	NCHC	6	6	Slave	Slave
Node_3	NCHC	6	6	Slave	Slave
Node_4	NCHC	6	6	Slave	Slave

O2: Compute & Store

B – 2 : Job Dispatcher (GridEngine and LSF)

Heterogeneous Computing Architecture

- We are developing the components of job dispatcher on top of two queuing systems, Grid Engine on VM group and LSF on Wind Rider



Proposed Components with GridEngine API

- drmaa_job_submit
  - drmaa_job_submit <"your_job [your_parameters]"> ["SGE parameters"]
  - drmaa_job_submit “/bin/date” “-p 1024”
- drmaa_job_status (return value: job id)
  - drmaa_job_status <your_job_id>
- drmaa_job_control
  - drmaa_job_control <your_job_id> <DRMAA_CONTROL_TERMINATE | DRMAA_CONTROL_SUSPEND | DRMAA_CONTROL_RESUME>
- Job dependencies:
  - drmaa_job_submit <"your_job [your_parameters]"> [“-hold_jid your_job_id"]

Proposed Components with LSF API

- lsfb_job_submit(returned value: job id)
  - Submit your job to LSF queuing system
  - lsfb_job_submit <"your_job [your_parameters]">
- lsfb_job_status
  - Get job status from job_id
  - lsfb_job_status <your_job_id>
- lsfb_job_control
  - To send stop, kill, resume signal to submitted job
  - lsfb_job_control <your_job_id> <SIGSTOP | SIGKILL | SIGCONT>
- Job dependencies:
  - To run a job upon the completion of specific job
  - lsfb_job_submit_jd <"your_job"> < "the job_id you want to wait for" >

Job status code

NARLabs

Code	Meaning
JOB_STAT_PEND = 0x01	job was pending
JOB_STAT_PSUSP = 0x02	Pending job was suspended
JOB_STAT_RUN = 0x04	job is running
JOB_STAT_SSUSP = 0x08	Running job was suspended due to overload
JOB_STAT_USUSP = 0x10	Running job was suspended by owner
JOB_STAT_EXIT = 0x20	Terminated with error
JOB_STAT_DONE = 0x40	Terminated without error
JOB_STAT_PDONE = 0x80	Post job was done successfully

NARLabs

Summary of major database tables and their physical size

Table Name	Row count	Physical Size	Note
fish_detection	1445.41M	322.26G	Abstracted information of detected objects in each frame
fish_species	663.93M	24.67G	Correlated of fish object to species catalog
fish	124.28M	21.01G	Abstracted information of detected fish objects
traj_species	97.29M	3.58G	Correlated tracking trajectory to species catalog
frame_class	11.61M	2.65G	Classification of video quality detailed to frames
fish_species_cert	32.55M	1.29G	Summary of detection/recognition certainty
summary_camera_39	7.13M	1.24G	Aggregation of information on camera id
summary_camera_46	7.12M	1.24G	
summary_camera_38	6.31M	1.10G	
summary_camera_37	4.46M	0.78G	
summary_camera_42	4.31M	0.75G	
summary_camera_44	1.49M	0.26G	
summary_camera_43	0.83M	0.15G	
video	0.63M	0.14G	Records of raw videos
processed_videos	0.78M	0.12G	Records of progress of video processing
summary_camera_41	0.63M	0.11G	
summary_camera_40	0.28M	0.05G	
video_class	0.53M	0.04G	Classification of video quality

NARLabs

Status of SQL database

- Database server is stable and running 24x7
- Able to support massive processing:
  - More than 2000 computing processes query to database, constantly having more than 100 processes doing insertion at same time into a huge table with ~1.5×10^9 rows.

NARLabs

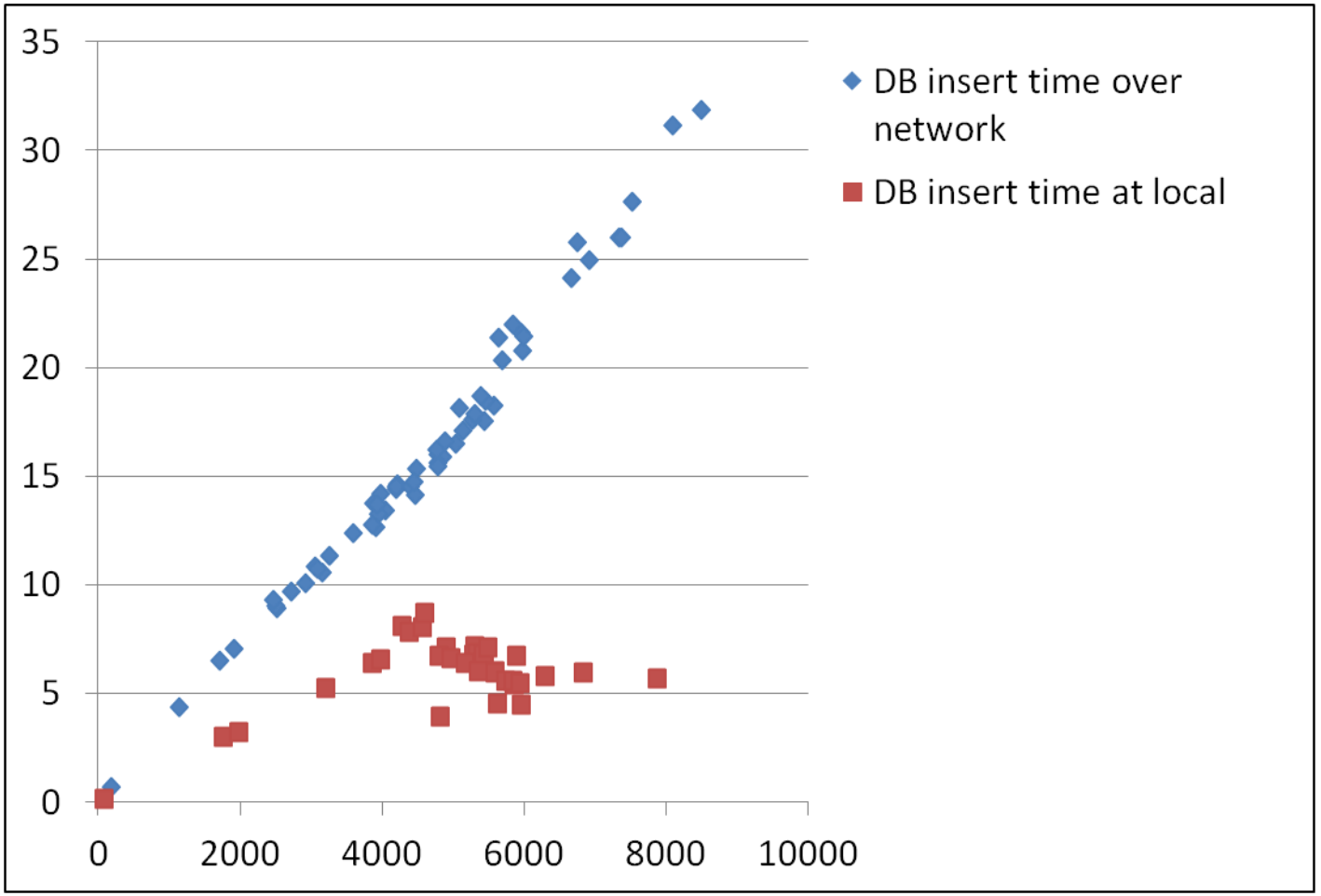
SQL database Performance Tuning

Statistic report from database server log

```
__ Questions ___________________________________________________________
Total       162.57M   95.5/s
 Com_       375.09M   220.3/s  %Total: 230.73
 -Unknown   370.01M   217.3/s        227.60
 DMS        157.32M   92.4/s         96.77
 COM_QUIT   163.37k    0.1/s          0.10
Slow 5 s    827.36k    0.5/s          0.51  %DMS:  0.53  Log:  ON
DMS         157.32M   92.4/s         96.77
 SELECT      56.40M   33.1/s         34.69       35.85
 INSERT      52.60M   30.9/s         32.36       33.44
 UPDATE      48.23M   28.3/s         29.67       30.66
 REPLACE     81.83k    0.0/s          0.05        0.05
 DELETE       1.21k    0.0/s          0.00        0.00
Com_        375.09M   220.3/s        230.73
 stmt_prepar 106.37M   62.5/s         65.43
 stmt_execut 106.37M   62.5/s         65.43
 stmt_close  106.28M   62.4/s         65.38
```

Too many communications!!

Time (second)



Data records

NARLabs

Bottleneck and solutions

- **Data insertion to huge table is time consuming**

- solution: packed insertions into one bundle and then send to database server once instead of sending them one by one.

ex:

```
insert into TABLE (a,b,c,d,e,f) values
(aa,bb,cc,dd,ee,ff),(aaa,bbb,ccc,ddd,eee,fff),(...),(...)
```

instead of

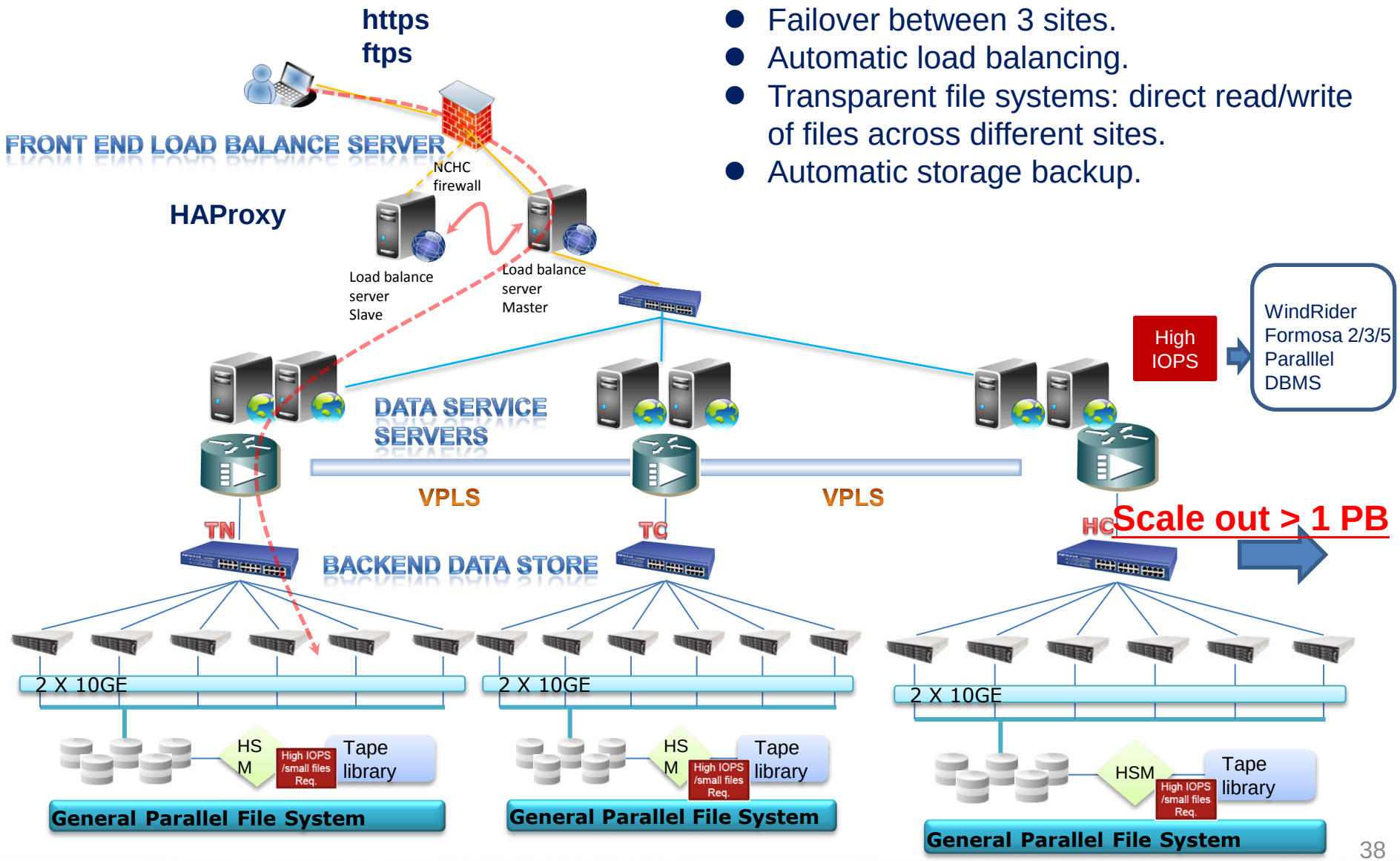
```
insert into TABLE (a,b,c,d,e,f) values
(aa,bb,cc,dd,ee,ff)
insert into TABLE (a,b,c,d,e,f) values
(aaa,bbb,ccc,ddd,eee,fff)
…..
```

- Disk IO and Network Latency
  - Migration of database server to a machine with larger capacity and lower network latency to HPC server
  - Move data store to SAN disk. Gain 1GB/s write, and 192MB/s read performance result from the move, and dramatically boosted efficiency of the detection processes that writes results into database heavily.
  - 4th July up and running, no deadlock happened while few thousands of detection processes were busy running. Made possible of finishing detection and classification on all videos.

Conclusions

- Provide tera-scale video data and sustainable infrastructure of compute, store, network and services to the F4K system with quality.
- Innovative data-intensive heterogeneous and distributed data infrastructure.
  - Multi-stage Data Streaming technology enables scalability and local intelligent for underwater monitoring systems.
  - Hybrid analytics platform: “Cloud” (VMs) + “Supercomputer”.
  - Hierarchical memory cache method speed up query performance up to 100x
  - Enable big table operations over ~1.5×10^9 records and achieve the tera-scale analytics through the hybrid analytics platform.
- Further impact: reference the system to build peta-scale data analytics platform.

Peta-scale data system for Earth Science Knowledgebase



- Failover between 3 sites.
- Automatic load balancing.
- Transparent file systems: direct read/write of files across different sites.
- Automatic storage backup.