

The slide features a decorative arrangement of six circles. In the top row, there are three circles: the leftmost is an outline, the middle is solid light purple, and the rightmost is solid light purple. In the bottom row, there are three circles: the leftmost is solid light purple, the middle is solid light purple, and the rightmost is an outline. The text is centered between these two rows.

# U-LITE: a proposal for scientific computing at LNGS

S. Parlati, P. Spinnato, S. Stalio  
LNGS 13 Sep. 2011

# 20 years of Scientific Computing at LNGS

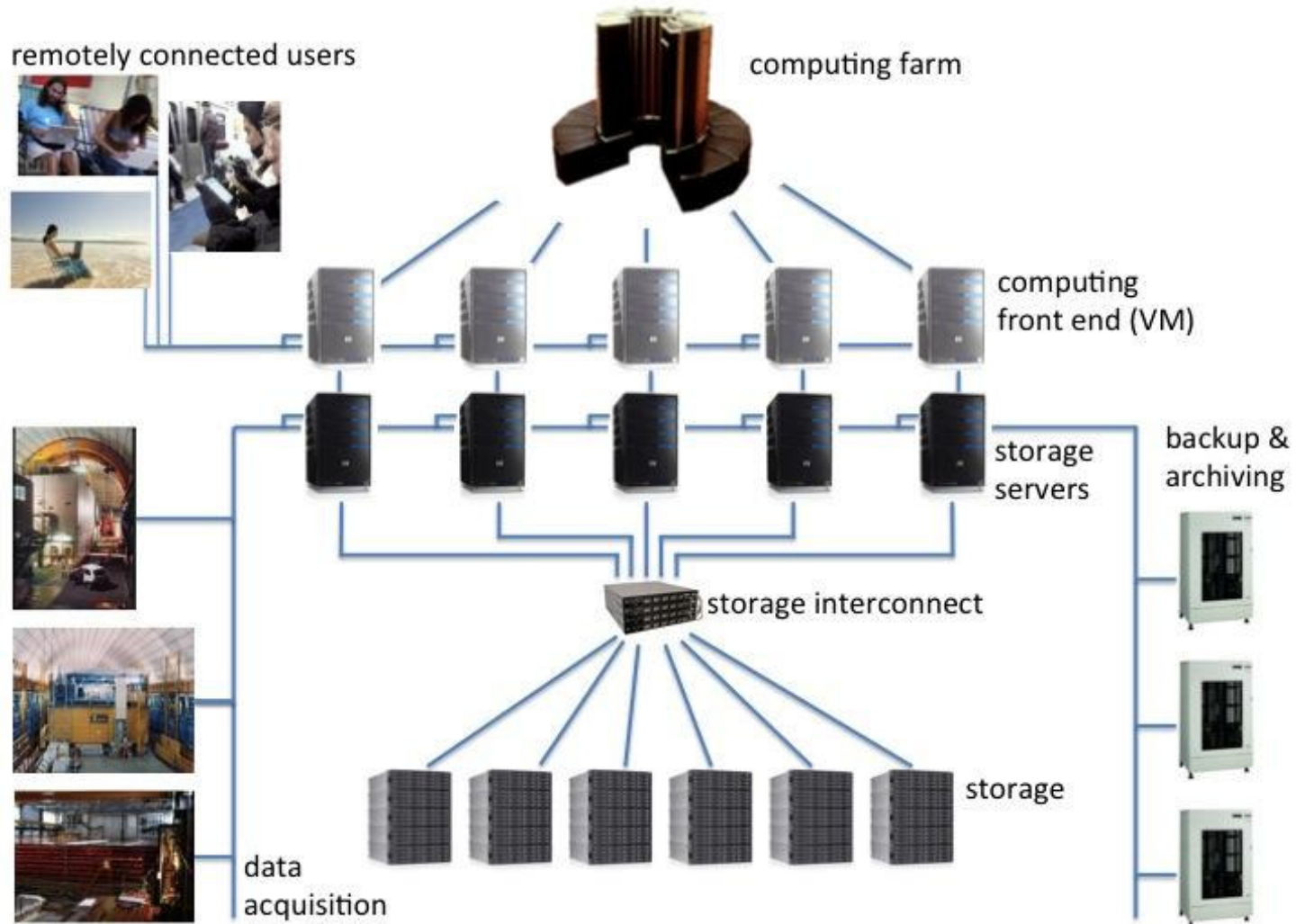
- Early '90s: highly **centralized** structure based on VMS cluster and DECNET network.
- Late '90s – 2000s: different unix systems (Digital Unix, Linuxes..); **heterogeneous** computing environments among collaborations. CNS continued to offer a complete computational environment (interactive, batch, storage, backup) even though various collaborations preferred to have their own farm.
  - 2008: first investigation on feasibility of a common infrastructure requested by INFN;
- 2010-2011: **shared** model, centrally managed; new techniques (virtualization) allow setup of heterogeneous environments on the same shared hardware -> U-LITE: right time to offer this service to future experiments!



# Why do we want to do it?

- Offer a better service to the Experiments: create a complete computing environment in the place where it is most needed: **where the data is actually collected**. (Our size is typically much smaller than CERN LHC experiments and there is no need of unlimited resources).
- Economic: resource sharing reduces expenses for infrastructure and optimizes resource exploitation
- Human resources: central management by **on-site experienced personnel** will allow experiments to save on human resources and avoid problems caused by unqualified or unmotivated personnel.

# Overview





# The U-LITE ingredients

- **Storage servers:** 1 or 2 for each experiment. They will copy data from DAQ to storage system, run first analysis or data reduction, save data to backup systems and present data to login servers and to the computing farm.
- **Computing front-end:** 1 or more per each experiment. They are accessed by users for sw development, for interactive tasks and to submit jobs to the computing cluster.
- **Computing farm:** many VM for each experiment, hosted inside real server (VM container). They will run batch jobs.
- **Authentication and Authorization:** users are managed centrally through Kerberos and LDAP databases.

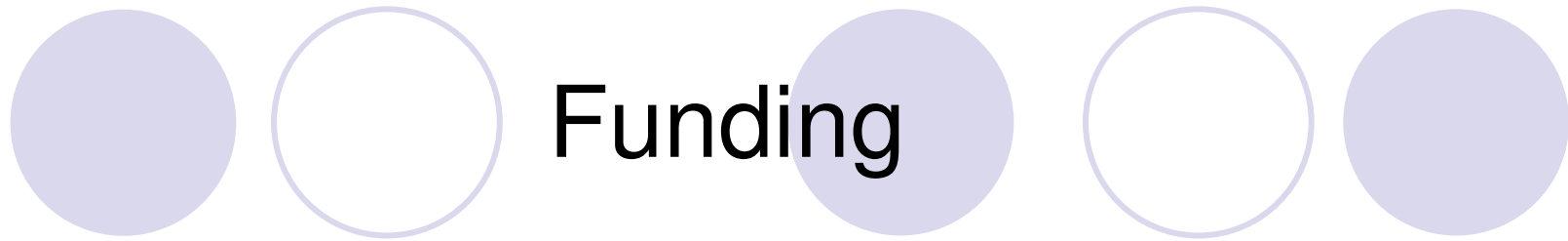
# Why experiments should join U-LITE?

- **Less work** for the experiments to setup a computational environment
- **Freedom** for the experiments to develop software for simulations and analysis on almost every linux dialect and with any kind of library and tool
- No need to adapt to GRID environment
- Stable and **continuous presence** of skilled staff on-site for infrastructure management
- **Savings** guaranteed by resource sharing
- Central monitoring of all the IT components



# Why can we do it?

- The CNS staff at LNGS has a **valuable experience** in managing all the U-LITE subsystems: it manages the network infrastructure and all the basic network services, multiple storage systems for a total of 150TB, two tape libraries, NFS and AFS servers, interactive login servers, LSF batch jobs system and centralized Authentication and Authorization systems.
- All the subsystems operate in **high availability** mode and are checked by a central monitoring system.
- The CNS has a long experience working with host **virtualization** (used for network services high availability).



- In the shared model we propose, funding should come from:
  - Experiments
  - Laboratory through the CNS
  - INFN National Computing committee.
- The way the three players will contribute has not yet been finalized and discussions with the INFN scientific and computing committee are necessary in the near future.





# Who is involved?

- Responsibilities

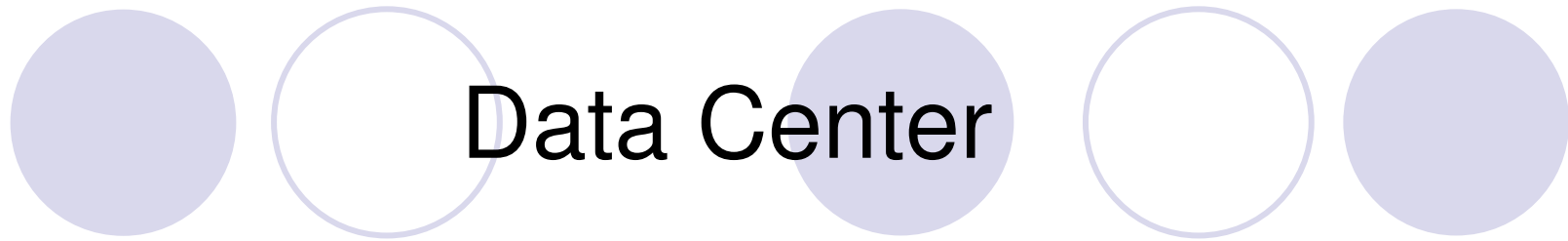
- Management and coordination: S. Parlati
- Technical and operational responsibilities: P. Spinnato and S. Stalio
- Supervision and planning: G. Di Carlo

- Technical staff

- Basic services: all the CNS
- Dedicated staff: P. Spinnato and S. Stalio
- **Human resources** requirements:
  - **3 FTE**, distributed among the existing staff and trainees, for the setup of the U-LITE project
  - **1.5 FTE** distributed between two or more members of the CNS staff when U-LITE will be up and running

# Current status and future work

- All the U-LITE HW and SW components are **tested and ready**: a first production environment has been operating since April 2011.
- HW rearrangement will be done in Sep/Oct.
- Storage and backup are **consolidated**; computing environment is young and under continuous development.
- The whole system is **scalable**
- The HW pool will increase as new experiments join U-LITE.

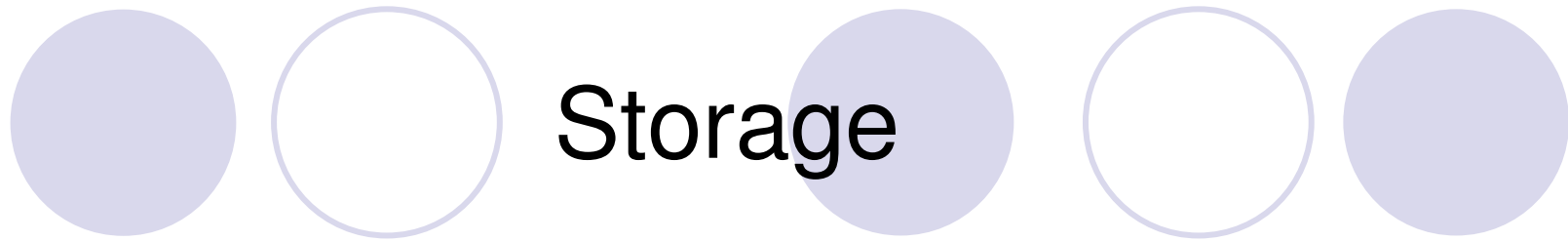


## Data Center

- All U-LITE subsystems live in the LNGS computing centre.
- The LNGS computing centre includes the main computer room and a second, smaller room in a different building.
- As for U-LITE, all systems are installed in the main computer room except for one of the two backup systems (server + tape library).

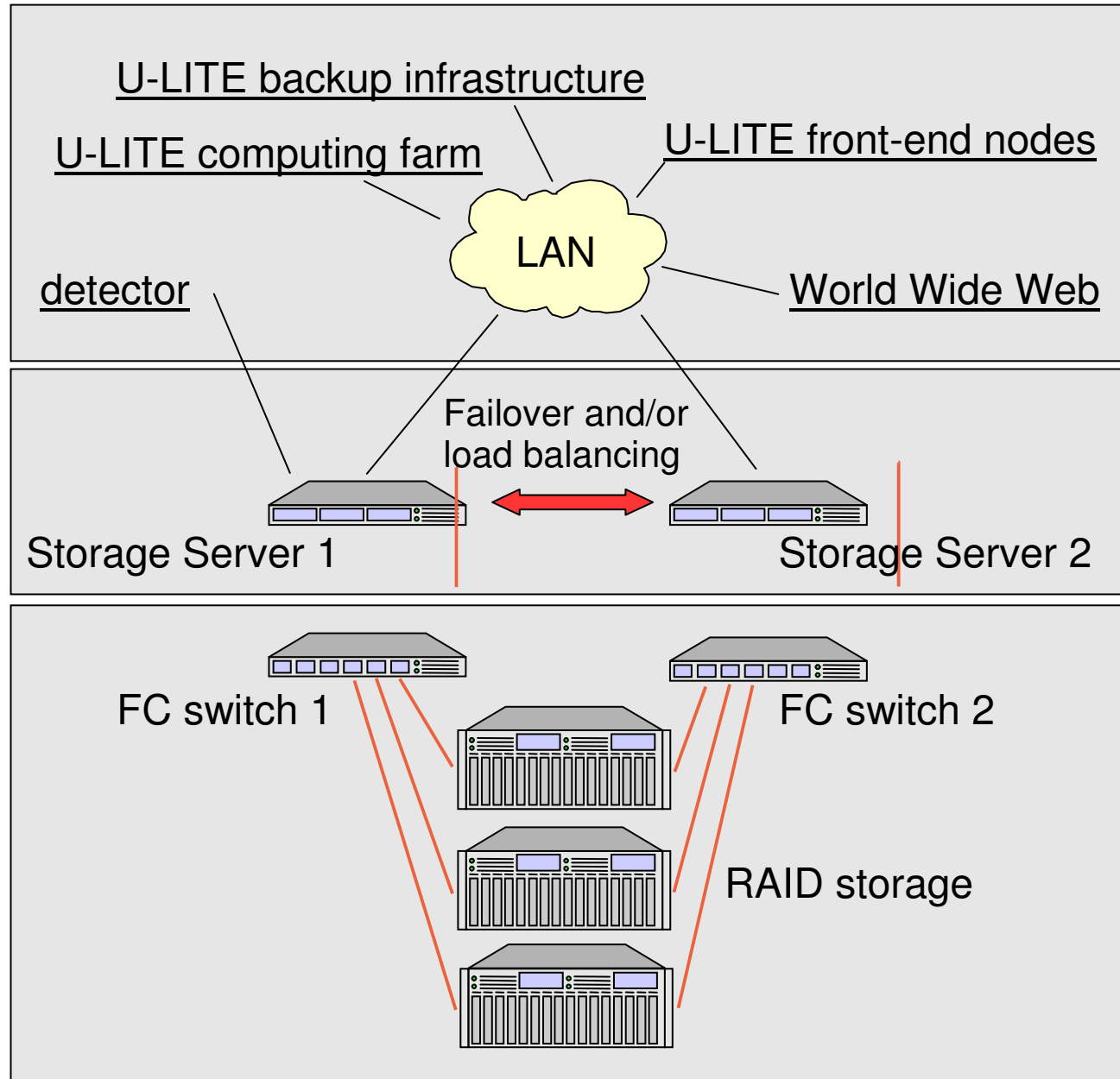
# Computer Room Characteristics

- Reliable power distribution infrastructure:
  - Diesel generator activates in case of black-out
  - 3 UPS in parallel cover short power outages
  - Redundant power distribution lines (2012)
- Multiple, independent, cooling systems. Failure of one system is not an issue for normal operativity.
- Computer room is large and near offices. Operator intervention is thus quick and easy.

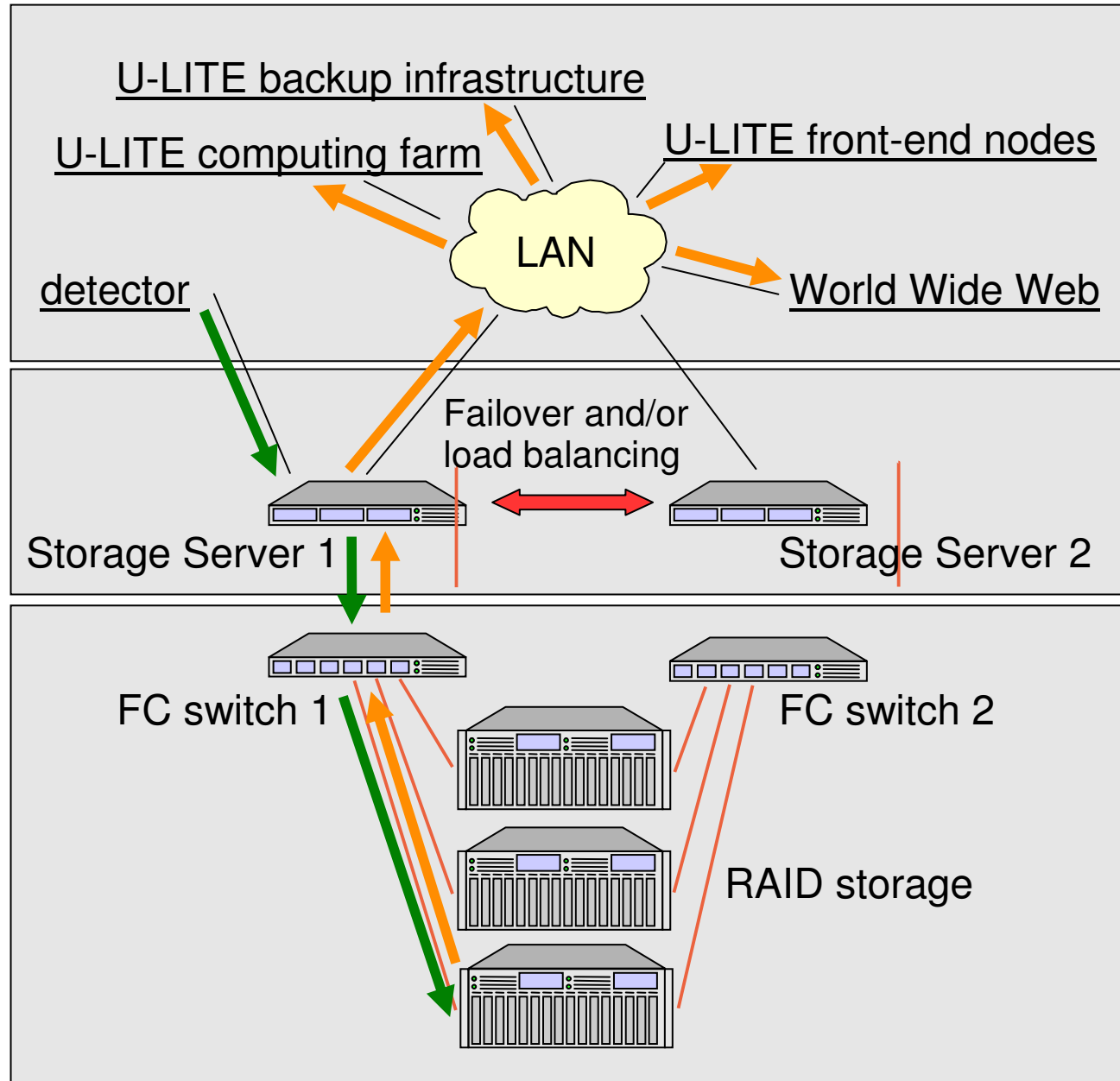


- Experimental data is copied from the detector DAQ to the U-LITE storage infrastructure via the LNGS LAN or a dedicated connection.
- Data is kept on reliable storage systems with redundant components (RAID, FC/iSCSI controllers, power supplies).
- Each experiment should have two storage servers connecting to the storage systems via different FC (or iSCSI) channels.
- Storage servers are managed by experimental collaborations. The LNGS CNS gives advice and guidelines.

*LNGS network infrastructure*



*LNGS network infrastructure*



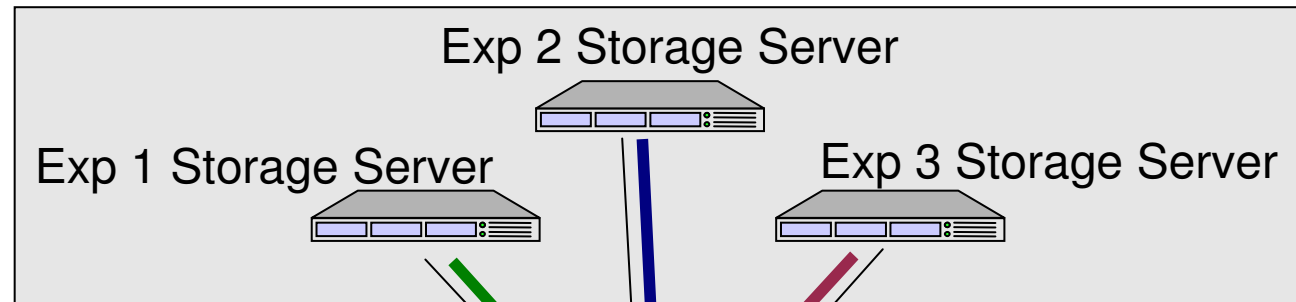


# Backup and Archiving

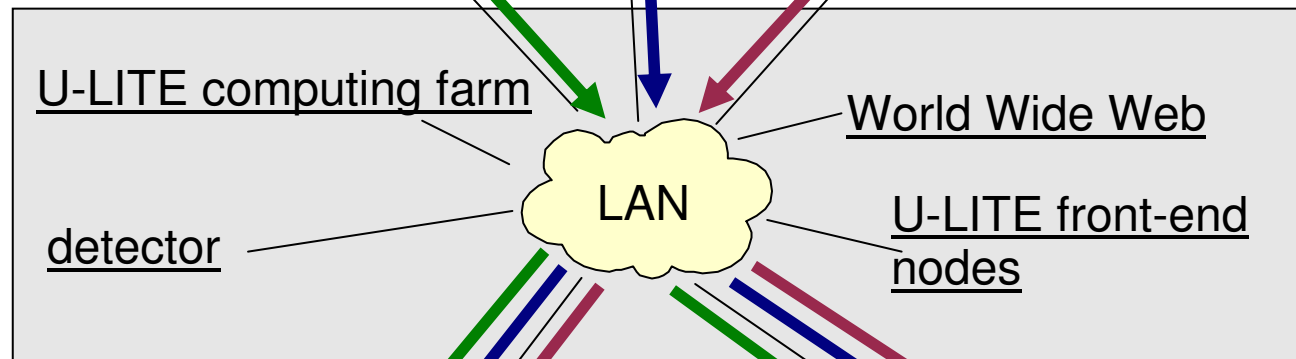
- Experimental data is backed-up on two separate tape libraries, located in separate buildings. Each data set is backed-up on both libraries.
- An open-source software (bacula) is used for data backup. As the data format is open, long term data readability is ensured. Data can be easily read on standalone systems outside LNGS.
- No periodic full backups are performed. Extremely large data sets (>10TB) may not be kept on-line inside the tape library.
- Long term data archiving in a dedicated room is available.
- Data backup and archiving is managed by the LNGS CNS. Storage policies may have to be agreed upon with experiments.



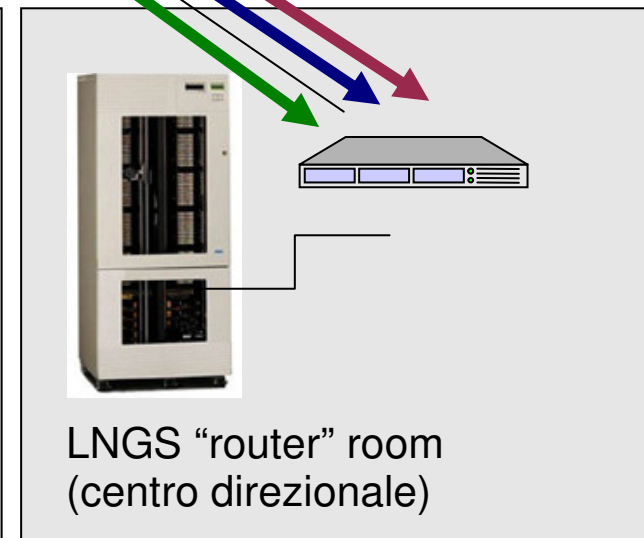
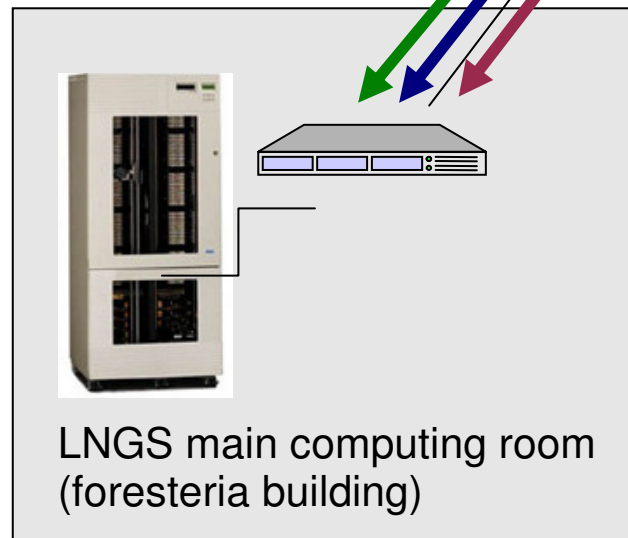
*Experiment-managed storage servers*



*LNGS network infrastructure*



*U-LITE backup infrastructure*





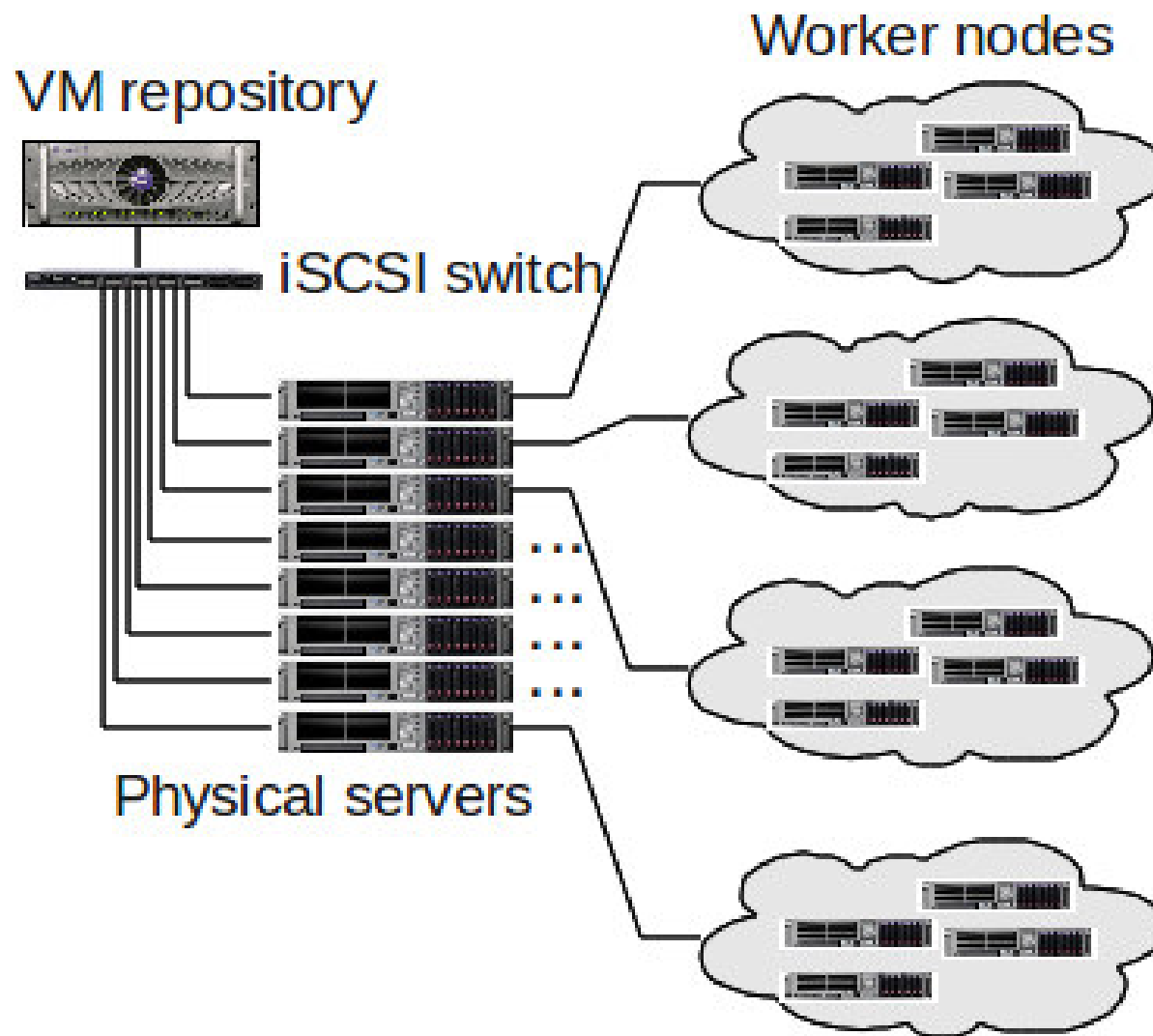
# Computing Cluster

- The U-LITE computing cluster is based on (KVM) virtual hosts as computing nodes. U-LITE computing nodes are hosted inside a cluster of physical multicore servers managed with the open source “Proxmox VE” software.
- Each experiment/workgroup has its own computing nodes based on a custom template.
- Computing nodes are automatically instantiated, migrated or turned off according to resource requests.



# Computing Cluster

- Computing node images are kept on a storage system that the physical servers share.
- Each computing node can always be started on or migrated to any server with no need to physically copy disk images.





# Monitoring

- Each component of the U-LITE system is continuously monitored for failures and overload.
- Mail/SMS messages are sent to CNS and/or to collaboration staff in case of problems.
- Historical data can be used for system tuning and for the discovery of weaknesses and bottlenecks.

# The U-LITE computing infrastructure

- Hardware made up of multicore servers, which act as containers of Virtual Machines
- Computing nodes are Virtual Machines, managed by a common user-transparent virtualisation system
- Job submission through front-end nodes using the open source TORQUE-Maui software
- Hardware financed by collaborations according to their needs, plus extra resources provided by CNS
- Priority allocation within reasonable time of financed resources enforced by a specifically designed allocation algorithm



# Why Virtual Machines

- Optimal exploitation of modern computing hardware calls for Virtual Machines
- By using VMs, collaborations are free to build their own computing environment, while using a common infrastructure
- CPU overhead is negligible, network and disk access overhead are acceptable



# Computing node build-up

- VM templates are developed by collaborations, CNS offers advice and support, validates template, makes computing node clones. Basic templates are also available.
- Front-end nodes (which may be template clones themselves) are used for software development, running interactive jobs, submitting batch jobs
- Authentication and Authorization are based on the LNGS AA infrastructure



A decorative header consisting of five circles in a row. The first, third, and fifth circles are solid light purple. The second and fourth circles are white with a light purple outline. The text 'Batch job submission' is centered over the second and third circles.

# Batch job submission

User connects to collaboration front-end node

User submits job on the most appropriate queue using TORQUE-Maui

PAN algorithm decides job priority

CRM wakes up VMs according to PAN scheduling

TORQUE-Maui allocates jobs to VMs

→ High availability of front-end nodes and TORQUE-Maui server is guaranteed by stand-by backup nodes



# Resource allocation in U-LITE

- Resource allocation must:
  - guarantee for each collaboration fast access to resources which it has paid for
  - maximise exploitation of available resources
  - distribute homogeneously free resources to all collaborations needing them
- We have developed a *Privileged Allocation of computing Nodes* (PAN) algorithm to fulfill such requirements

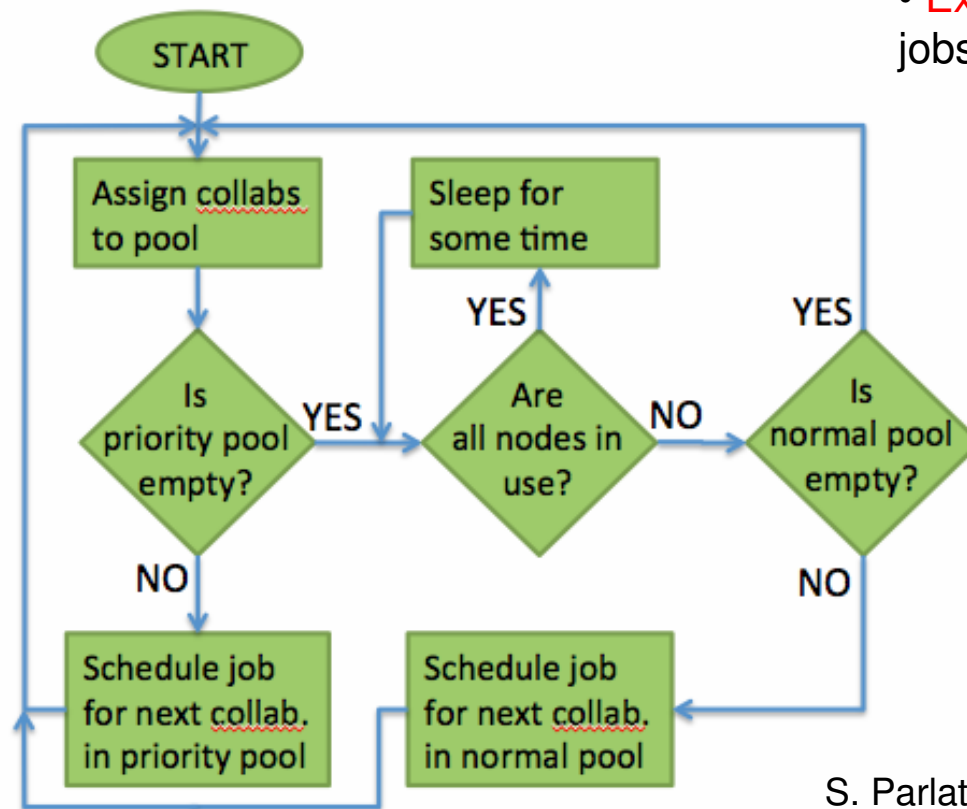
# The PAN Algorithm

2 pools:

- **Priority pool** for collabs that have not exhausted its resource quota
- **Normal pool** for collabs that have exhausted its resource quota and need extra resources

3 queues:

- **Long queue** for unlimited duration jobs, jobs can only run on collab. resource quota
- **Short queue** for limited duration jobs, job can run everywhere
- **Express queue** for very urgent and short jobs, jobs can run everywhere

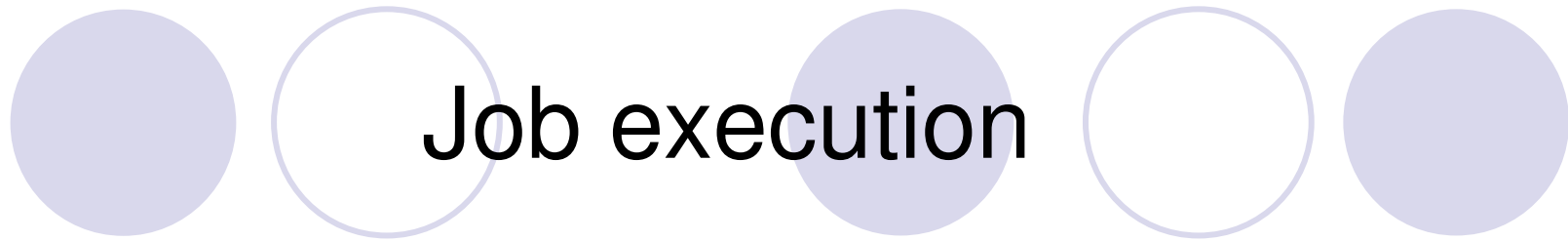




# Resource allocation

future developments

- Hot migration to slow nodes in case Privileged Allocation requires it
  - VM disk space resident in a common Storage Area makes this straightforward
  - Queue model changes accordingly: from duration-based to allocation-based
- VM hibernation in case all resources are occupied
- VMs used also to run interactive jobs
- Hardware configurable at job submission
- Resource allocation mechanisms highly configurable and open to replacements



- Process runs on VM computing node, username is the one who submitted job on front-end node
- Input data accessed via network
- After job completion, VM stays up in case other jobs requiring it arrive, or switches off, if needed, to free resources.

### CRM Monitor

User	Running jobs	Max slots
DEFAULT	0	10
stallo	1	10

Jobs run on Mon 12 September 2011

Queued	2
Run	2
Completed	1
Deleted	0

Jobs run on Wed 31 August 2011

Queued	8
Run	7
Completed	7
Deleted	6

Jobs run on Mon 29 August 2011

Queued	1
Run	1
Completed	1
Deleted	0

Jobs run on Fri 12 August 2011

Queued	5
Run	5
Completed	5
Deleted	0

Jobs run on Thu 11 August 2011

Queued	1
Run	1
Completed	1
Deleted	0

Server running on host qmaster  
 Start time Tue Aug 9 14:28:15 2011  
 Local time Mon Sep 12 09:44:48 2011

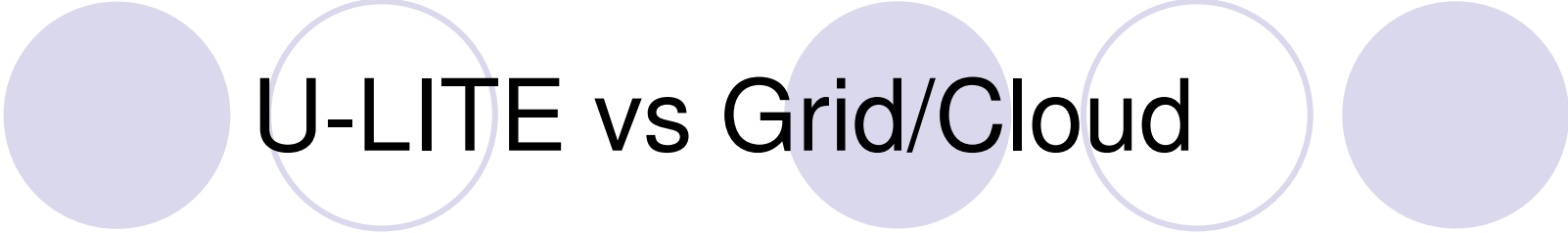
Torque is running	09/12/2011 09:43:58;0100;PBS_Server;Job:482.qmaster.lngs.infn.it;dequeuing from gs-long, state COMPLETE
Maui is running	09/12 09:44:24 INFO: scheduling complete. sleeping 30 seconds
CRM is running	Mon Sep 12 09:44:39 2011 - jobs (queued/running): 0/1, nodes (down/up): 2/10, free resources (slots/cores/RAM): 12/6/20438
CRM last log entry	Mon Sep 12 09:43:58 2011 - jobs (queued/running): 0/1, nodes (down/up): 2/10, free resources (slots/cores/RAM): 12/6/20438

Node	RAM	CPUS	Server	Last Op	Idle	State	Properties	Job	Run Time	Queue	Owner	Group
ge-login	4096	2	hnode00	none	931640	free	gerdanode,ge-login,always_on					
vnnode001	4096	4	hnode00	none	2918810	free	teonode,vnnode001,always_on					
vnnode002	4096	4	hnode03	none	2918810	free	teonode,vnnode002,always_on					
vnnode003	4096	4	hnode01	none	2918810	free	teonode,vnnode003,always_on					
vnnode004	4096	4	hnode01	none	2918810	free	teonode,vnnode004,always_on					
vnnode005	2048	2	hnode03	migrate	2737350	offline	gerdanode,vnnode005					
vnnode006	2048	2	hnode04	none	2918810	offline	gerdanode,vnnode006					
vnnode007	2048	2	hnode02	none	50	offline	gsnode,vnnode007					
vnnode008	2048	2	hnode02	start	0	offline,job-exclusive	gsnode,vnnode008	483.qmaster	00:06:29	gs-long	stallo	200
vnnode009	2048	1	hnode04	none	0	down	lvdnode,vnnode009					
vnnode010	2048	1	hnode04	none	0	down	lvdnode,vnnode010					
vnnode011	2048	2	hnode03	none	2918810	offline	gerdanode,vnnode011					

Server	Slots (used/free/total)	CPU cores (used/free/total)	RAM (used/free/total)	CPU speed
hnode00	2/2/4	6/2/8	8192/26/8166	2327.369
hnode01	2/2/4	8/0/8	8192/26/8166	2327.938
hnode02	3/5/8	5/3/8	5120/11280/16400	1595.785
hnode03	3/1/4	8/0/8	8192/26/8166	2327.453
hnode04	2/2/4	3/1/4	3072/9236/12308	2327.414

Queue	Priority	Max Time	Group	Nodes	Jobs (running/queued/total)
gerda-long	200	9999:00:00	gerda	gerdanode	0/0/0
gerda-short	100	02:00:00	gerda	gerdanode	0/0/0
gerda-xpress	300	00:20:00	gerda	gerdanode	0/0/0
gs-long	200	9999:00:00	gs	gsnode	1/0/1
gs-short	100	02:00:00	gs	gsnode	0/0/0
gs-xpress	300	00:20:00	gs	gsnode	0/0/0
lvd-long	200	9999:00:00	lvd	lvdnode	0/0/0
lvd-short	100	02:00:00	lvd	lvdnode	0/0/0
lvd-xpress	300	00:20:00	lvd	lvdnode	0/0/0
teo-long	200	9999:00:00	teo	teonode	0/0/0
teo-short	100	02:00:00	teo	teonode	0/0/0
teo-xpress	300	00:20:00	teo	teonode	0/0/0

# Job monitoring



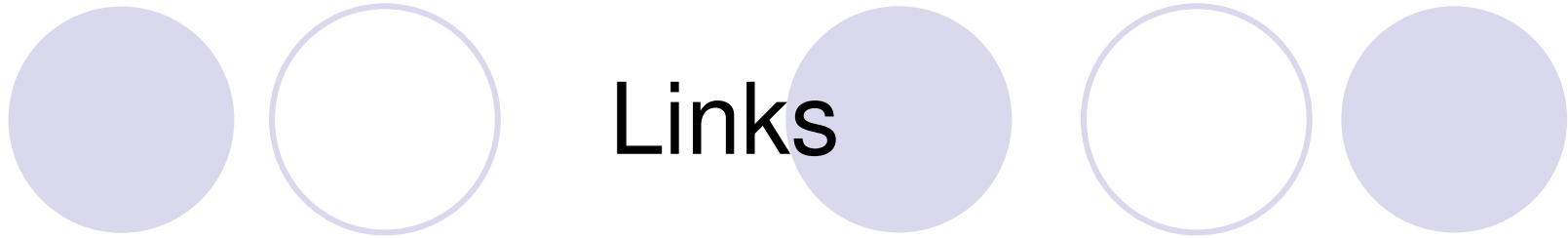
# U-LITE vs Grid/Cloud

## Pros

- System managers are in direct contact with users (in contrast to Cloud)
- Extremely user-friendly: IT infrastructure is transparent to user (in contrast to Grid)
- Data and computing are co-located (in contrast to both)

## Cons

- (little) risk of resource saturation
- Maybe more costs for hardware



- <http://u-lite.lngs.infn.it>  
U-LITE main site
- <http://qmaster.lngs.infn.it>  
Job monitoring
- <http://computing.lngs.infn.it>  
LNGS Computing service main site