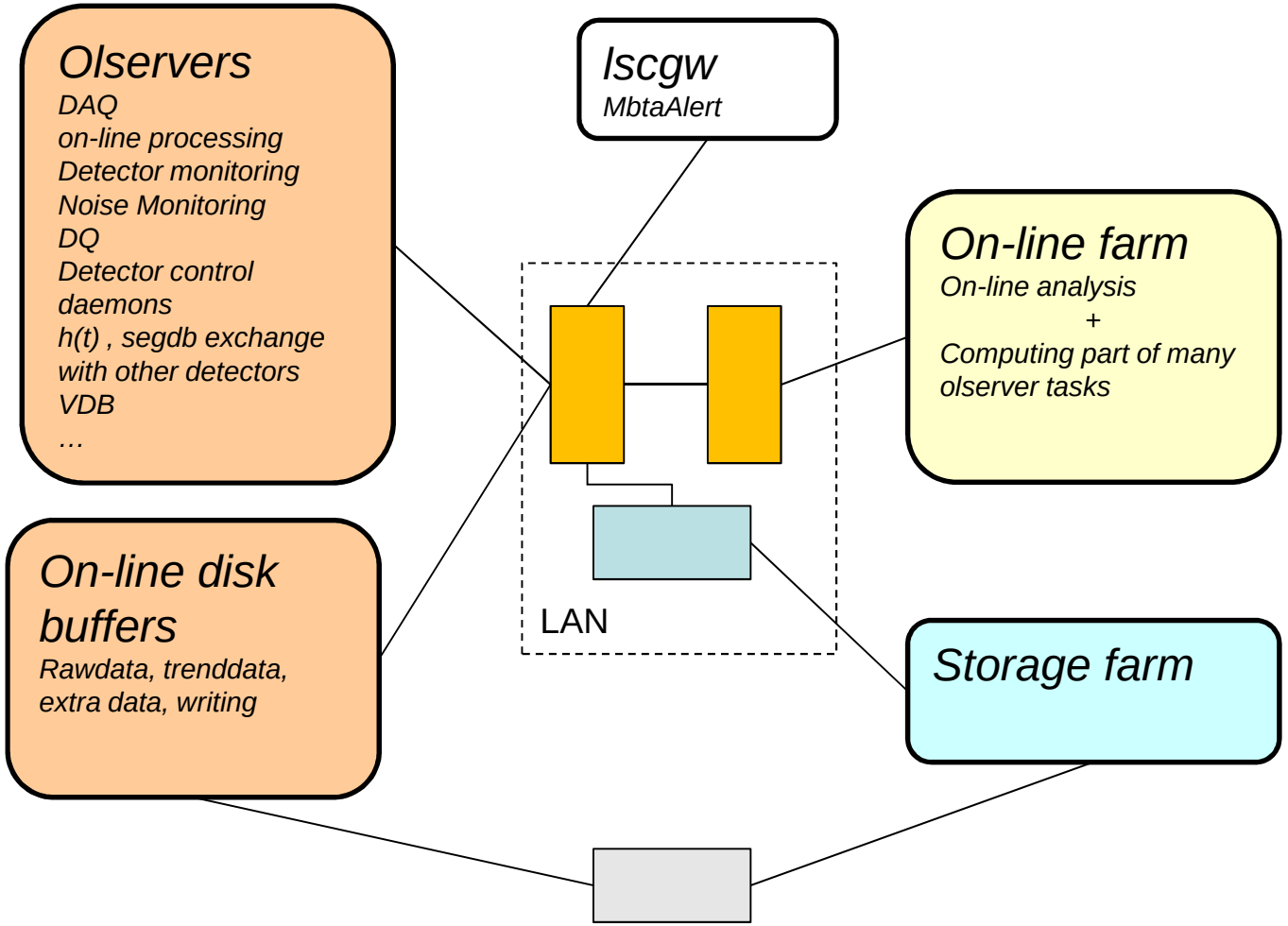
Issues for on-line computing platforms

Stefano Cortese : VDAS meeting 22-11-2012

Summary

- Current status: olservers and on-line farm
- Limitations
- Merging olservers / on-line farm ?
- Proposed reorganization
- Actions needed from Virgo
- Test system

Current on-line computing schema



Current olservers: 20 nodes + 8 vguests , 96 core, 104GB RAM

Alignment Locking Injection DAQ Data Quality	Hosts the Cm framework daemons (Db, Cm ns…)	olserver	Intel(R) Xeon CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	DAQ, Alp	olserver1	Intel(R) Xeon CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	DAQ Frame builders: FbmAlp + FbmSt	olserver2	Intel(R) Xeon CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	h-reconstruction	olserver3	Intel(R) Xeon CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	DAQ frame builders: FbmTmp + Fbm50 + DQ BRMSMoniSM	olserver4	Intel(R) Xeon CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	spare	olserver5	Intel(R) Xeon CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	DAQ compress + resize	olserver9	2 x Intel(R) Xeon CPU,E5450 4x @ 3.00GHz, 8 GB RAM
	spare	olserver10	2 x Intel(R) Xeon CPU,E5450 4x @ 3.00GHz, 8 GB RAM
DAQ Data Quality Noise Monitoring	MonitoringWeb1 (Spectrograms)	olserver6	Intel(R) Xeon(R) CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	MonitoringWeb2 , vdqsegdb exchange with Ligo	olserver7	Intel(R) Xeon(R) CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	spare	olserver8	Intel(R) Xeon(R) CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	DQ reprocessing , D-NMAPI test	olserver15	intel(R) Xeon(R) CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	online DQ	olserver16	intel(R) Xeon(R) CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	Noise Monitoring (SpectroMoni)	olserver17	intel(R) Xeon(R) CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	Detector Monitoring (Monis)	olserver18	intel(R) Xeon(R) CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	spare	olservertest	intel(R) Xeon(R) CPU,E5450 4x @ 3.00GHz, 4 GB RAM
	~~dead~~	~~olserver11~~	~~2 x Intel(R) Xeon(TM) CPU 3.40GHz, 2 GB RAM~~
	spare	olserver12	2 x Intel(R) Xeon(TM) CPU 3.40GHz, 2 GB RAM
	MonitoringWeb Internal	olserver13	2 x Intel(R) Xeon(TM) CPU 3.40GHz, 2 GB RAM
	DQ reprocessing, vdqsegdb exchange with Ligo	olserver14	2 x Intel(R) Xeon(TM) CPU 3.40GHz, 2 GB RAM
Noise monitoring Slow control Suspension Control	WDF	olserver31	virtual guest 2 VCPU 2.4GHz, 4 GB RAM
	Coherence	olserver32	virtual guest 2 VCPU 2.4GHz, 4 GB RAM
	Slow control/DAQ IMMS	olserver33	virtual guest 2 VCPU 2.4GHz, 2 GB RAM
	Suspensions + VSU	olserver34	virtual guest 2 VCPU 2.4GHz, 2 GB RAM
	NOEMI	olserver35	virtual guest 2 VCPU 2.4GHz, 4 GB RAM
	Tangodb / Vacuum	olserver36	virtual guest 2 VCPU 2.4GHz, 2 GB RAM
	CAM Applications	olserver37	virtual guest 2 VCPU 2.4GHz, 2 GB RAM
	BacNet server (Eurotherm)	olserver38	virtual guest 2 VCPU 2.4GHz, 2 GB RAM
Any	VDB internal server	vdb73	2x Intel Xeon CPU,E5335 4x @ 2.00GHz, 4GB RAM

olnodes: 97 nodes 324 cores, 104GB RAM, 528GB RAM (includes 6 nodes down)

Opteron cores: 64 x248 2GHz + 260x275 2.2GHz

Current tasks assignment (from workarea):

What	Who	Olnode	Specification	Cores#	last update	note
		1		2		down
		2		**2**	**11/07/2012**	**free**
DQ performance	Florent	3	bi-pro/8GB	2	11/07/2012	
MBTA	Benoit, Frederique	4-19	bi-pro/8GB	32	11/07/2012	ADE low frequency/large templates require large memory cores
NoEMi	Alberto	20	bi-pro/8GB	2	11/07/2012	8GB is needed
WDF	Elena	21-22	bi-pro/8GB	4	11/07/2012	8 GB need not certain
		23-26	**bi-pro/8GB**	**16**	**11/07/2012**	**free**
		27	bi-pro/8GB			down
		28	bi-pro/8GB		12/08/2011	Having trouble. ECC memory pb
		29-30	**bi-pro/8GB**	**4**	**11/07/2012**	**free**
System test for part 1 of the farm	Administrator	31-32			08/09/2010	
MBTA	Benoit, Frederique	33-34	quadri/4GB	8	11/07/2012	
		35-51	**quadri/4GB**	**68**	**11/07/2012**	**free. could be used temporarily by MBTA**
MBTA	Benoit, Frederique	52	quadri/4GB	4	11/07/2012	
WDF	Elena	53-55	quadri/4GB	12	11/07/2012	
Omicron	Florent, Nicolas	56-66	quadri/4GB	44	11/07/2012	not used before aux. channels available
Spectrograms	Didier	67-70	quadri/4GB	16	11/07/2012	
DQ	Didier	71	quadri/4GB	4	12/07/2012	not used permanently
		70-86	**quadri/4GB**	**60**	**11/07/2012**	**free**
		87-89	**quadri/8GB**	**12**	**11/07/2012**	**free. 8 GB nodes**
		90-94	**quadri/4GB**	**24**	**11/07/2012**	**free**
		95	quadri/4GB	4	02/09/2010	down
system test for part 2 of farm	Administrator	96-97		8	08/09/2010	

Current Virgo+ architecture: on-line-farm / olservers

On-line farm

- Multinode on-line computing farm with loose access to network storage (via NFS)
- Physical nodes partitioned quasi-statically to science groups
- Activity includes both on-line/in-time computing during runs and development/analysis off runs

Olservers

- Same multinode architecture as on-line farm
- Original mission: to support DAQ tasks, on-line processing and Detector daemons and monitoring
- Main criticalities: insure ITF operation , provide the rawdata fluxes for the first writing

Merging of on-line-farm / olservers ?

Why decoupled systems?

1. Most critical system could be freezed *(olservers still SL4.5 vs. olnodes SL5.3) => good for operation*

   More complex or ever changing tasks/platforms => environment ever changing (reconfigurations/reallocations/upgrades/troubleshooting/security patches) => *good for data analysis*

2. Less impact of non-critical tasks versus critical ones (resources contention)

Does decoupling olservers/on-line-farm still hold?

Probably:

Yes for: rawdata assembling , ITF software infrastructure (daemons), operation monitoring + all tasks ending in writing non reproducible data?

Not for: all tasks with computing backend on on-line farm

What about: on-line data exchange with other detectors ?

on-line-farm current limitations

Resources allocation

Multinode farms: more suited to batch queuing systems or massively parallel tasks

In our usage:

- the tasks are not deterministic, include development, interactive use, manual processing => large peaks
- many resources (RAM, cores) are wasted because not shareable among tasks

Provisioning is difficult, requires hw changes (increase of RAM , change ethernet network, etc.) => slow

Performance

Many new tasks are disk-based => I/O to the storage farm is a bottleneck

Users workflow

Users expectation: general use (remote accessibility , graphical tools, general not-nearline data access usage, web servers, db servers) => not all met

Others ? : …

on-line-farm proposed model

Multiprocessor farm instead of multinode farm:

- Few many-cores nodes
- Allocation met through partitioning techniques: coarse-grain via virtualization, fine-grain via OS based resources partioning *(linux cgroups, containers)*

Generic HEP computing model instead of a specific Virgo one

- Performances enhanced both in network I/O and in network filesystems parallel access with a generic workload ,mix of sequential, random, concurrent access patterns (protocol integration with storage farm, pNFS over RDMA, DCB network/SAN convergent architecture)

Various users workflows

- IPv4/IPv6 public addresses for direct connection from outside (subject to rules..)
- Batch queuing system for a subset of machines *(others interested besides Noemi ?)*
- GRID access for a subset of machines /storage ?

Back to on-line farm / olservers roles

Limitations of (any) computing-farm model:

- Nodes are indifferent as seen by applications (except for sizing), cannot be hw bound
  => if specific non-shareable hw is needed (tolm, GPUs, …) , better a standalone system

Advantages of many-cores nodes for olservers still hold

- Other applications would benefit of direct access to shared memory ?
- Virtualization layer useful also for olservers critical tasks: optimization of provisioning, possible highly available redundant configurations

Previous considerations on decoupling for olservers running a subset of critical tasks still hold

Conclusions

Actions needed from Virgo:

1. Decide whether the proposed model for the on-line farm substitution could fit the scientific needs (meaning, it's not an obstacle to..)
2. Specify/add the workflows needed (BQS, on-line data exchange with other detectors, etc..)
3. Start the process of requirements collection from all the applications for the sizing estimate.
   Need of a standard template grasping application behaviour: interface with data, interface with other applications, sw requirements, etc..

EGO requirements: we need a small-scale test system

Proposal: install a minimum of 24core 64GB RAM machine replacing farmnxx when tests are finished

Note 1: not a production machine for ER runs

Note 2: phase 1 testing without 10Gb ethernet (2-4Gb bonding) , phase 2 testing with DCB chain (convergent switch+storage)