



**Joint Institute for Nuclear
Research**
SCIENCE BRINGING NATIONS
TOGETHER



NATIONAL RESEARCH
SOUTH URAL STATE
UNIVERSITY

**The International Conference - Mathematical Modeling and
Computational Physics 2017**

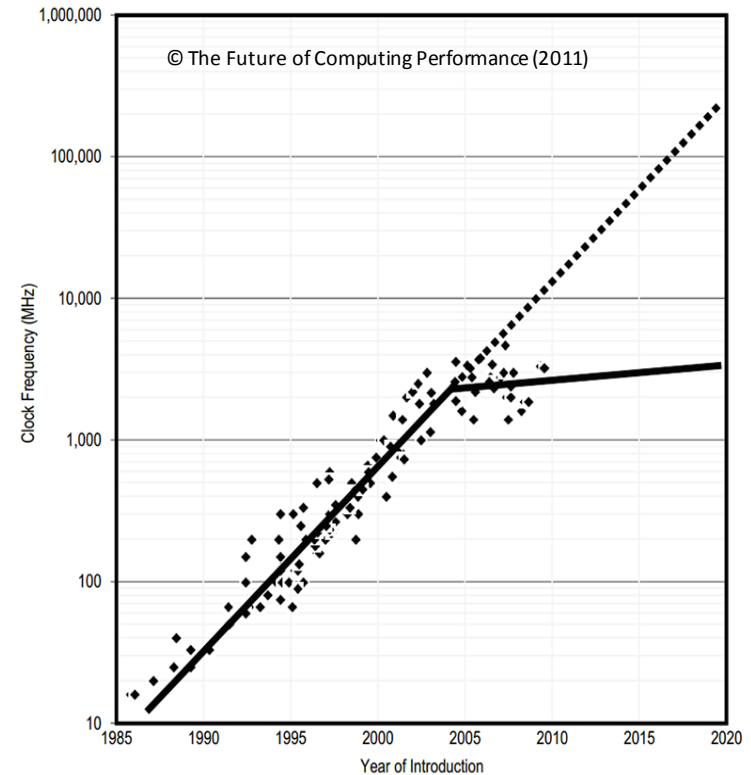
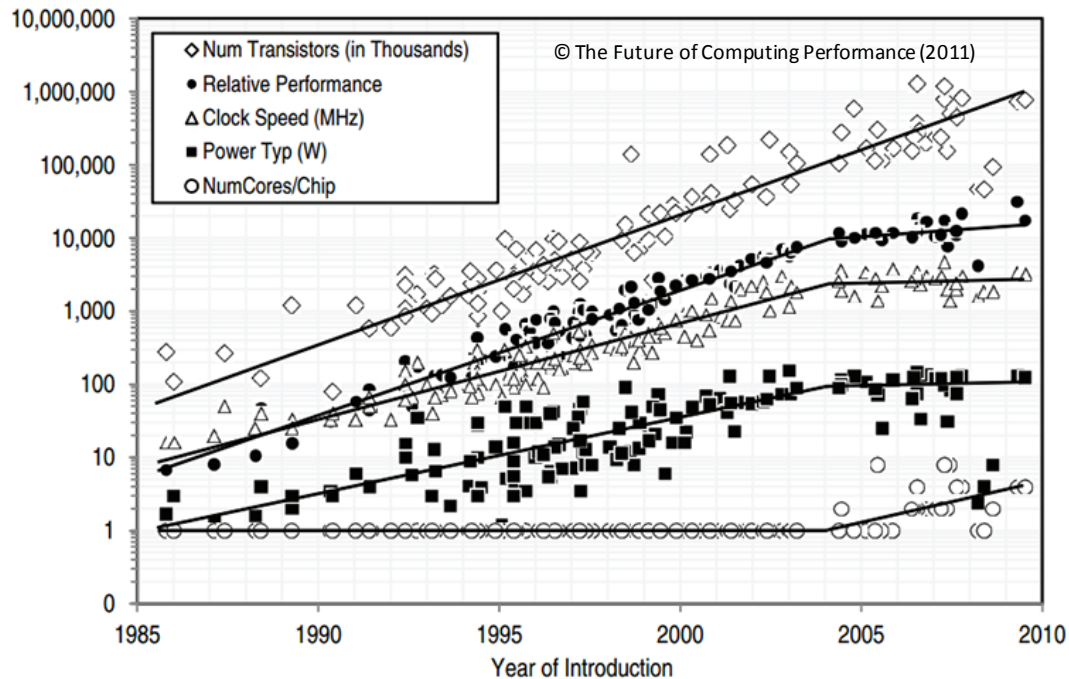
Application of SLURM, BOINC and GlusterFS as Software Complex for Sustainable Modeling and Data Analytics

Dr. Kaftannikov I.L., Mr. Kashansky V.V.

South Ural State University
Department of Electrical Engineering and Computer
Science

Dubna, 2017

Dynamics of the Hardware Computational Market



- Reaching "saturation" region of the clock frequency on current systems
- Software and hardware clustering of engineering solutions
- Operation in frames of centralized and decentralized clustering architectures

Birth of the System

- The problem of infrastructure deployment arose during 2015-2016;
- Processing of medium-sized datasets in frames of several semantic-networks projects and non-linear system modelling projects led by Dr. I.L. Kaftannikov;
- We faced the problem of bridging the gap between low-cost computing cluster and high-end computing system.

Market of GRID Computing Systems

- Globus Toolkit
- BOINC
- HTCondor
- gLite
- UNICORE

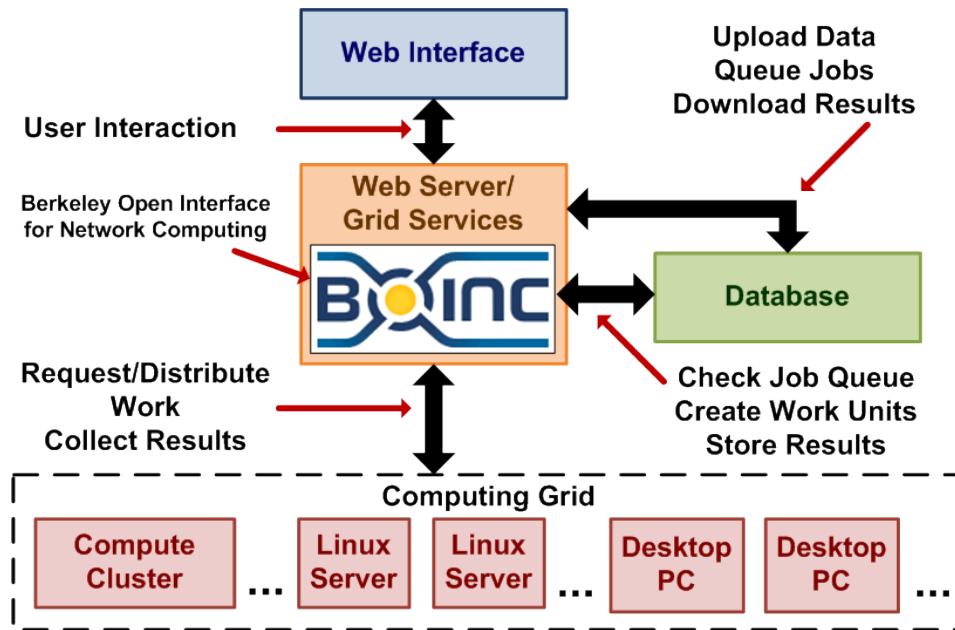


BOINC System Description



- **BOINC** (Berkeley Open Infrastructure for Network Computing) — free software (LGPL license) computing platform. Implementation of the “volunteer GRID computing” model;
- Modular centralized service-oriented (SOA) platform. Base components are implemented, using several programming languages, including C, C++, Python for OS Linux. Metadata is mostly stored in MySQL, there are also several configuration files. User-friendly web-interface is written in PHP (backend), HTML and JavaScript (frontend);
- Highly popular and internationally visible. There are many clients for different operation systems working on different hardware solutions.
- It is used by us for deployment of low-cost computational platform at early stages.

BOINC System Architecture in Context of the OS Linux



- Runs variety of daemons, behind the web server;
- Application and system layers are loosely-coupled;
- Almost all the components of the system are well-replaceable.

In the spirit of SOA - system administrator is able to control the whole system via web-interface, including: computational units, user base and host base.


Revealed Benefits and Disadvantages

- ✓ Well defined documentation; Completely free and open source, as well as working environment (OS);
- ✓ The system is widely known; There is also a cross platform client, running on the most OS;
- ✓ Active participants community; Easy to administer and control as well as deploy

- ✗ Incorporates file-server and web-server by default
- ✗ Uses MySQL for storing metadata by default
- ✗ Processing of queries is implemented via PHP and C/C++ CGI
- ✗ Inconsistent with local job distribution, which requires low transition rates and fast acting system
- ✗ The question with data storage is open and undefined as system operates at higher layer of abstraction

BOINC Performance

Data obtained during simulation tasks



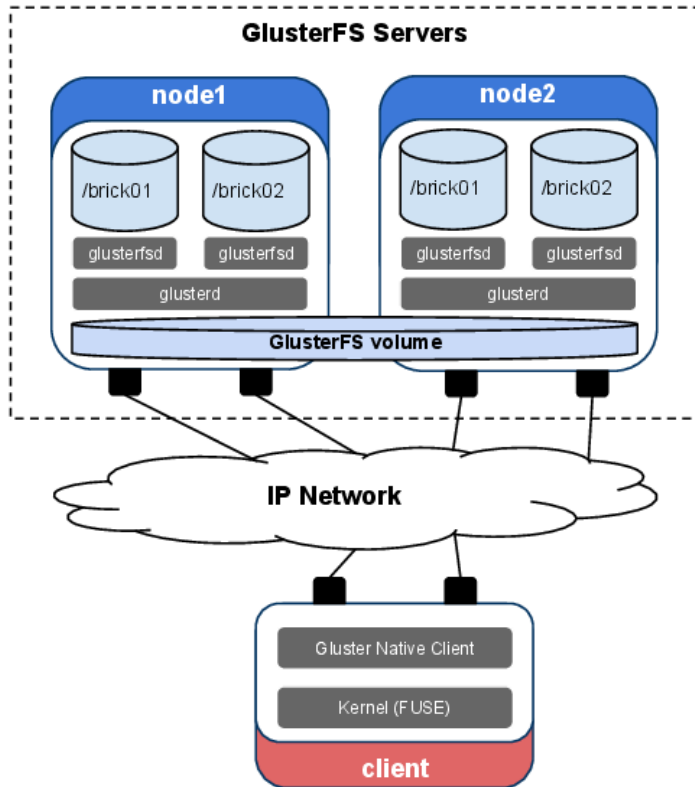
N	50	500	5000	20000	50000	100000	200000
T	1.464	14.428	147.029	578.367	ND	ND	ND
N	50	500	5000	20000	50000	100000	200000
T	31.369	61.959	74.736	141.396	481.966	963.932	1926.864
N	50	500	5000	20000	50000	100000	200000
T	ND	ND	ND	104.971	100.971	167.818	262.548

The results of calculations on the interval 0 to 10 at step 0.000025 seconds. N is the number of input vectors, T is the time calculated by the UNIX Time utility in seconds

BOINC Optimization. Architecture Redefinition

- Dividing of file-transferring (FTP) and web-servicing (HTTP, RPC);
- Optimization of MySQL, migration to MariaDB or Percona Server (XtraDB). Change of the DB schema. Possible transition to NoSQL;
- Transition from Apache+CGI to nginx+FastCGI. Kernel tuning for non-blocking processing;
- Changing of the core. Transition to HTCondor, UNICORE, Globus Toolkit;

GlusterFS



- GlusterFS is a distributed, parallel, linearly scalable file system.
- Using TCP/IP stack, GlusterFS can combine data stores on different servers into a single parallel network file system.
- Highly popular and internationally visible, currently developing by RedHat.
- It is used by us for deployment of low-cost highly robust storage;

GlusterFS Choice Highlights

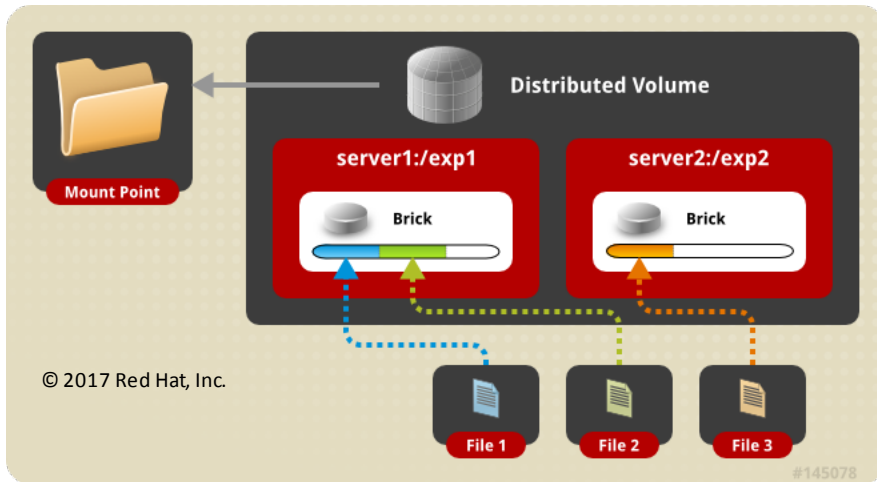
- Decentralized
- Scalable
- Low-dependable
- Easy to deploy and maintain
- In active development

	HDFS	iRODS	Ceph	GlusterFS	Lustre
Architecture	Centralized	Centralized	Distributed	Decentralized	Centralized
Naming	Index	Database	CRUSH	EHA	Index
API	CLI, FUSE REST, API	CLI, FUSE API	FUSE, mount REST	FUSE, mount	FUSE
Fault detection	Fully connect.	P2P	Fully connect.	Detected	Manually
System availability	No failover	No failover	High	High	Failover
Data availability	Replication	Replication	Replication	RAID-like	No
Placement strategy	Auto	Manual	Auto	Manual	No
Replication	Async.	Sync.	Sync.	Sync.	RAID-like
Cache consistency	WORM, lease	Lock	Lock	No	Lock
Load balancing	Auto	Manual	Manual	Manual	No

Input/Output	HDFS		iRODS		Ceph		GlusterFS		MooseFS	
	I	O	I	O	I	O	I	O	I	O
1 × 20GB	407s	401s	520s	500s	419s	382s	341s	403s	448s	385s
2 × 20GB	626s	422s	1070s	468s	873s	495s	426s	385s	504s	478s
1000 × 1MB	72s	17s	86s	23s	76s	21s	59s	18s	68s	4s
2 × 1000 × 1MB	96s	17s	179s	20s	85s	23s	86s	17s	89s	4s

Performance analytics made by Depardon, Benjamin, Gaël Le Mahec, and Cyril Séguin. In the paper "Analysis of six distributed file systems." (2013): 44

Fitting of GlusterFS

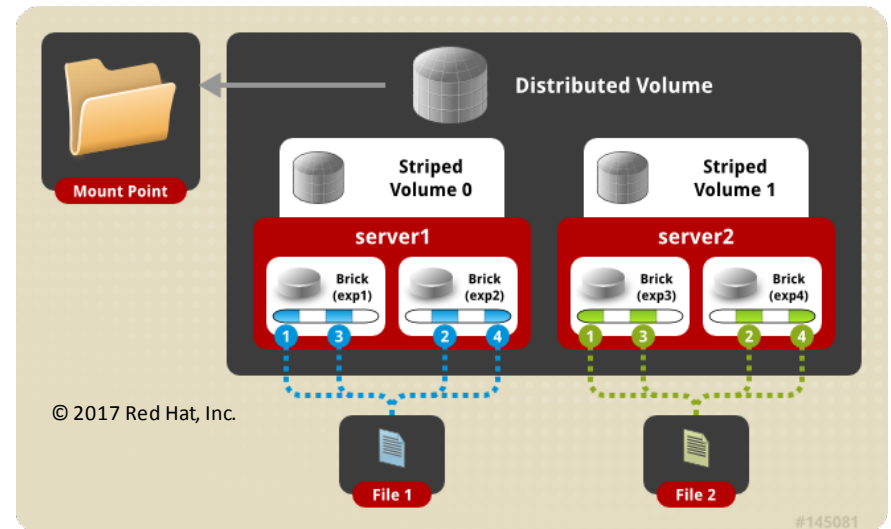


- GlusterFS has a client and server component;
- Servers are typically deployed as storage bricks, running a glusterfsd daemon to export a local file system as a volume;
- The glusterfs client process, which connects to servers with a custom protocol over TCP/IP, InfiniBand or Sockets Direct Protocol.

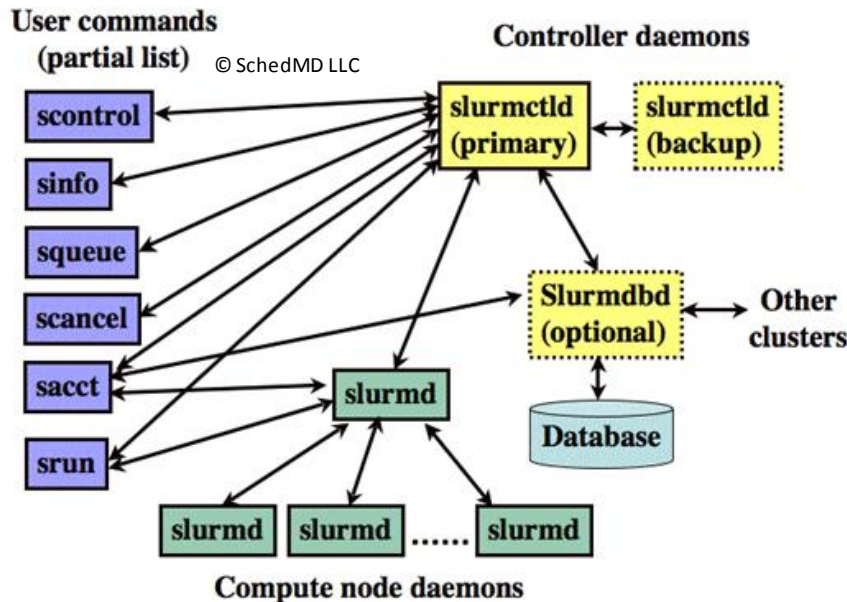
By bringing GlusterFS to our BOINC stack we change the storage mechanisms, allowing store our data in next way:

- Distributed
- Replicated
- Stripped

Any combinations of the modes above are also possible.

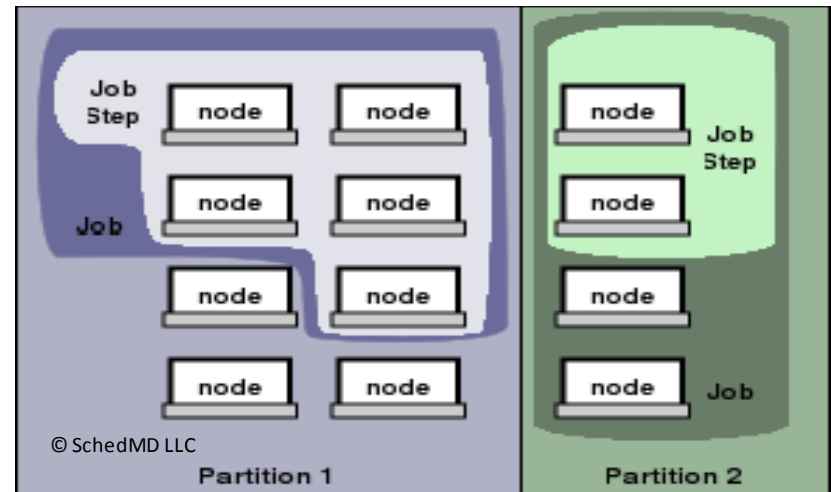


Architecture of the SLURM scheduler



Simple Linux Utility for Resource Management (SLURM) - is an open, reliable and well-scalable cluster resource management system with a task scheduler, used for both large and small Linux clusters.

SLURM consists of a `slurmd` service running on each compute node and a central `slurmctld` service running on the controlling node (optionally with a backup copy of the controlling node).



Choosing Distribution Model and Algorithms

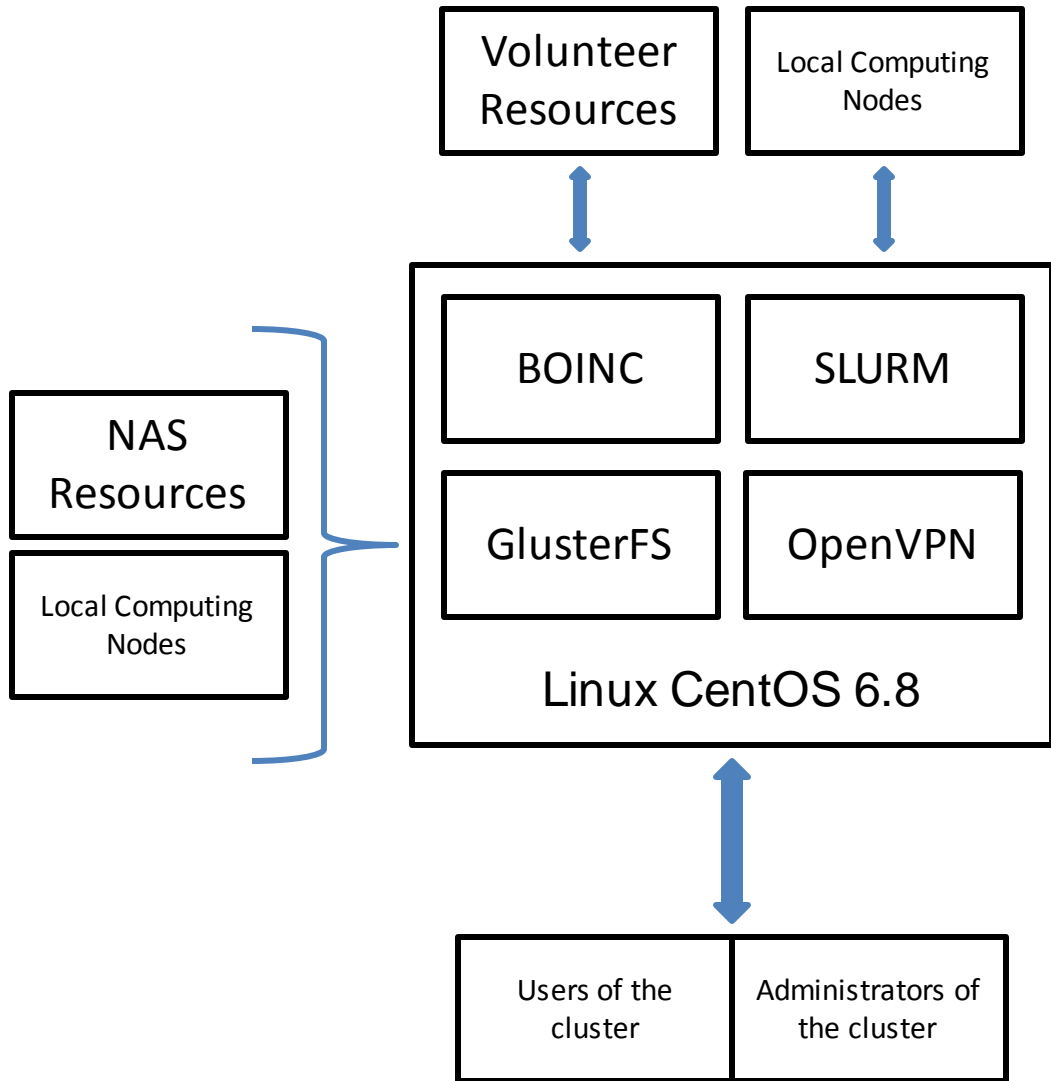
- Define the class on time domain;
- Long-performing (value order of days, hours) tasks are running on the volunteering clusters;
- Long-performing jobs with intensive data exchange are running on the local cluster and scheduled via SLURM;
- Local fast-performing (value order of minutes, hours) jobs are running on BOINC or on the local cluster;

Hardening of GlusterFS and SLURM.

Overlay networks

- Incorporating local high-performance nodes and subclusters;
 - Running the storage and other managing software inside an overlay network to increase security level;
 - Easiness of joining new nodes to a network;
- ✓ Implemented via OpenVPN software;

Synergy of Systems and Final Stack



- BOINC and SLURM manage computational jobs;
- Interaction with users and administrators is via SSH;
- OpenVPN provides overlay for inner communication;

Conclusion. Ways to Grow

Using of SLURM, BOINC and GlusterFS as a platform for our computations allowed:

- Get a horizontally scalable computing system with minimal costs for organization and subsequent maintenance;
- Solve the idleness problem with the inner main resource pool of the department;
- Department to socialize its research, allowing outside researchers to offer their own algorithms (resources) and run them on our system;
- After the implementation of the stack, it was possible to process much more data (due to GlusterFS), automate functions related to the dissemination, collection and aggregation of results, the creation of reporting data for different type of jobs.

Thank you!