IJCSS International Journal of Complex Systems in Science

Int. J. Complex Systems in Science
vol. $\mathbf{1(2)}$ (2011), pp. 89–95

Application workflow deployments on computing grids with a complex network topology

F. Prieto-Castrillo[1,†], A. Astillero[2], E. Rey-Espada[1],
M. Boton-Fernández[1] and M. Arsuaga-Rios[1]

[1] *CETA-CIEMAT, Research Centre for Energy, Environment and Technology*

[2] *Dept. of Computer and Communications Technology, University of Extremadura*

Abstract. Preliminary results for application workflow deployments over a simplified GRID with a complex, non-trivial communications network topology and a homogeneous computing capacity are addressed. Application workflows are modeled as Directed Acyclic Graphs (DAG) where communication costs are assumed to depend on the underlying network *hopcount*. A simple DAG model is implemented to deduce the applications's overall *makespan* over a minimal GRID consisting on schedulers and computing resources uniformly deployed over the network. The obtained network topology is built from an exhaustive set of active measurements on the Spanish e-Science GRID infrastructure. This results into a power-law degree distribution with a characteristic exponent of $\alpha = 2.87$. Optimization criteria suggest that lower hierarchization schemes, where the distinction between scheduling and computing nodes vanishes, are preferable to the existing design.

Keywords: Complex-Networks, Small-World-Graphs, Distributed-Computing, GRID-Computing
MSC 2000: 05C82, 68Q85

[†] **Corresponding author:** francisco.prieto@ciemat.es
Received: October 10, 2011
Published: October 24, 2011

1. Distributed computing and GRID environments over complex networks

Computing GRIDs can be thought as a set of software and hardware services deployed over a communications network resulting into a single entity [1]. The GRID lies somewhere in between the centralized workload management inherited from the distributed clustering paradigm and the *Peer-to-Peer* Computing (P2P) -fully decentralized- scheme. In this regard, active research in being developed for the implementation of a *self-organized* GRID [2], [3],[4]. In a minimalist GRID (i.e. no other interacting GRID services such as resource discovery or data replication are considered), the workload management problem is highly tied to the interaction of two service types: scheduling and processing. These are usually deployed on a dedicated high-speed research network such as the Spanish Academic and Research Network (RedIRIS) [5]. The overall picture is a distributed system where server nodes (brokers) dispatch GRID tasks (jobs) to computing clusters or Computing Elements (CEs) spread across different administrative and geographic domains.

Although GRID technology has already been intensively exploited -specially in the high energy physics community- the GRID is still considered as a prototype [6]. A challenging question still unsolved is which is the optimal number of computing nodes in a GRID environment [7]. Also, the influence of network topological features in the efficiency of a grid system is an open problem [8], [9].

This work presents an optimization scheme for scientific application workflow scheduling in computational GRIDS with non-trivial topologies. In particular, the average shortest-path length of the underlying network along with the optimal broker/CE proportion are related with time performance metrics. Although the assumed topology was built from real data, other communications network metrics (e.g. latency, bandwidth or background traffic) are neglected.

2. Allocating tasks into the computing infrastructure

Scientific applications can be modeled through Abstract Directed Acyclic Graphs (ADAG) consisting on the 3-tuple: $ADAG \equiv (G_J, \mathbb{M}, \mathbb{O})$, where $G_J = (J, E_J)$ is a directed graph with a jobs vertex set J, a tasks size map (in bytes) $\mathbb{M}$ and an estimated transfer files map between adjacent jobs $\mathbb{O}$.

On the other hand, the communications network is assumed as an undirected graph $G_R = (R, E_R)$ with a communication nodes set R consisting on routers, switches or even Autonomous Systems (AS).

As stated, GRID workload is modeled by the random deployment of both brokers (B) and CEs (C) sets on G_R through random variables β y θ respectively.

Each Broker-CE pair from the available $n_B = |B|$ brokers and $n_C = |C|$ CEs constitutes a *computing mode* $S \equiv B \otimes C$.
Every job is mapped to computing modes through the map: $\mathbb{A} : J \rightarrow S$ (Fig.1 (left)). This mapping is assumed as an stochastic process $\{\mathbb{A}(J_k) : J_k \in J\}$.

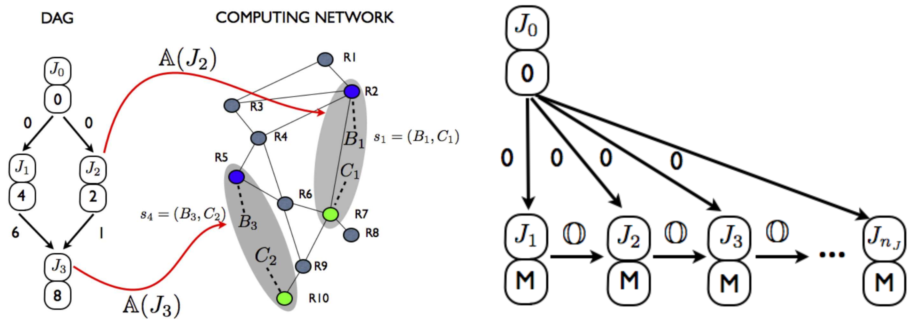

Figure 1: Allocation process from an arbitrary DAG into the computing network (left) and assumed ADAG model for the present study (right).

Then a *Concrete DAG* (CDAG) can be obtained from the former ADAG by specifying network communication costs $\mathbb{T}^c$ and computation capacities $\mathbb{T}^c$: $CDAG = (G_J, \mathbb{T}^s, \mathbb{T}^c)$
A widely used application performance metric is the Critical Path Length (CPL) [10]; the longest path in the weighted DAG. In the present case its ensemble mean $< CPL >$ from the allocation process is used as main performance metric.

In this work a simplified ADAG model is assumed where the first ω jobs are mapped to the same mode (*clusterization*) while the remaining $n_J - \omega$ are uniformly distributed among the entire mode set $n_S = n_B \cdot n_C$.
As shown in Fig.1 (right), from an entry node J_0 (with both zero size and transmission cost) $n_J = |J|$ jobs are forked. Each child job J_i has equal M size (bytes) and carries a fixed communication cost $\mathbb{O}$ (s).
A uniform GRID resource mapping is also assumed: $P(\beta = R_j) = P(\theta = R_j) = 1/n_R, \forall R_j \in R$. Furthermore, every CE holds the same processing speed p (bps) and broker's are assumed as simply task dispatchers (i.e. scheduling algorithms, resource discovery services or other inner logic are neglected).
The underlying communications network has been assumed as homogeneous both in latency and bandwidth. Furthermore, paths are assumed as geodesic, symmetrical and persistent. Finally, no packet fragmentation or background traffic effects are considered. Under these assumptions, communication costs are proportional to the path *hopcount*. Although this is clearly unrealistic for

a real GRID over the internet, for latency-bounded applications (i.e. small file transfer sizes so that bandwidth heterogeneity is a secondary effect) over dedicated high bandwidth networks these limits seem reasonable as a first approach to reveal topological effects.

By focussing also in cases where $n_R \gg 1$ y $n_S \gg 1$ the $< CPL >$ (Eq.1) renders:

$$< CPL >= n_J \left(T^q + \frac{M}{p} \right) + (n_J - \omega)(4\eta \bar{l})[1 + \left(\frac{1}{4n_C^3} \right) \Gamma] \tag{1}$$

where T^q (s) is the CE queuing time, $\bar{l}$ is the mean shortest path length, η (s) is the proportionality parameter between the communications costs and the path *hopcount*, and the communications overhead parameter Γ is given by:

$$\Gamma = \frac{1 - z\gamma_C}{z} \tag{2}$$

where the *centralization factor* $z \equiv n_B/n_C$ and $\gamma = n_C/n_R$ have been introduced.

3. Communications network tomography

The infrastructure topology has been built from the e-Science National GRID Initiative Infrastructure (NGI) [12] which is deployed over the Spanish Academic Research Network (RedIris) [5]. Sets of small sized *probe jobs* where sent to the every NGI CE with *traceroute* commands to every other CE in the generic Virtual Organization [1] iber.vo.ibergrid.eu. The whole map was then recomposed as a picture of the topology *seen* by a grid job (see Fig.2 (left)). In the tests, the selected routes where always persistent over the whole experiment. It must be highlighted, though, that since *traceroute* metrics do not always allow inter-domain resolution, some nodes in the resulting network where in fact whole Autonomous Systems (AS) composed by several communication nodes. Hence, the reported topology lies somewhere in between the router and the AS level [13]. This may explain the obtained exponent $\alpha = 2.87 \pm 0.13$ in the corresponding power-law degree distribution (Fig.2 (right)). The obtained metrics for NGI topology under iber.vo.ibergrid.eu VO with $n_R = 189$ nodes, 477 edges , $n_B = 5$ brokers and $n_C = 26$ CEs, where: average degree $\bar{k} = 5.03$, average clustering coefficient $\bar{cc} = 0.105$ and mean shortest path length $\bar{l} = 4.24$.

Since the obtained empirical topology resulted into a small-medium network, to cope with the resulting *poor statistics*, the methods addressed in [11] where implemented.

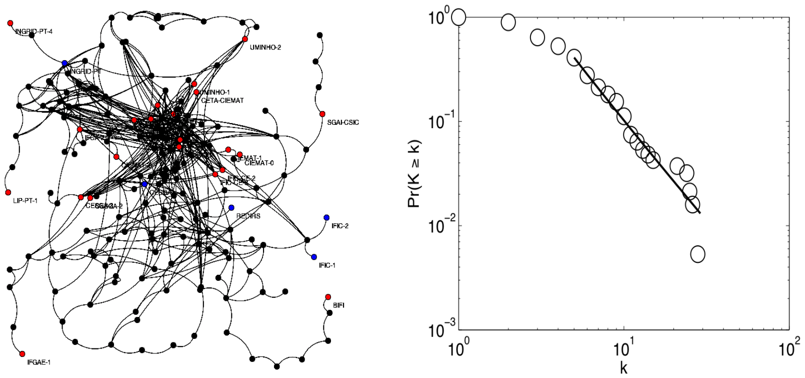

Figure 2: Experimental topology found for the NGI infrastructure (left) and power-law degree distribution found with the methods described in [11] (right). Also the least squares fit for $P(K \geq k) \approx k^{-\alpha}$ with $\alpha = 2.87 \pm 0.13$ is shown

4. Conclusions

Through simplifying assumptions analytical expressions of performance metrics for application deployment on complex-network based GRIDS where derived. As expected, $< CPL >$ decreases as $\omega \rightarrow n_J$ (i.e. total clusterization). This effect is likely due to the finite CEs processing capacity that has been implicitly assumed. This is clearly unrealistic. In a more accurate approach, the finite processing capacity should result in a balancing term in $< CPL >$ which increases with ω.

It is also noticed the linear dependence of the communications overhead with the mean shortest path length $\bar{l}$. A global optimum for $\Gamma = 0$ is achieved at $n_B = n_R$.

If $n_B \ll n_R$, the maximum optimization achievable carries a factor of $\frac{1}{4n_C^3}$. This factor is likely to be decreased for more efficient (i.e. intelligent) designs, such as a betweenness or degree based GRID resources mapping. By increasing the computing resources number and since $0 \leq \gamma_C \leq 1$, the maximum optimization can only be achieved for finite P2P networks when $\gamma_C = 1$ (com-

Γ	Max. centralizations ($n_B = 1$)	P2P ($n_B = n_C$)
Finite net ($n_R < \infty$)	$n_C - \gamma_C$	$1 - \gamma_C$
Infinite net ($n_R \rightarrow \infty$)	n_C	1

Table 1: Communications overhead Γ factor in four limit cases

puting nodes approach the network's communication node number). Results are summarized in Table 1.

Acknowledgements

CETA-CIEMAT acknowledges the support received from the European Regional Development Fund through its Operational Programme 'Knowledge-based Economy'

References

[1] Foster et al. The anatomy of the grid: Enabling scalable virtual organizations. International Journal of High Performance Computing Applications (2001) vol. 15 (3) pp. 200

[2] M. Boton-Fernandez, F. Prieto-Castrillo, M. A. Vega-Rodriguez. Self-adaptive deployment of parametric sweep applications through a complex networks perspective. The 12th International Conference on Computational Science and Its Applications, Part II, LNCS 6783, pp. 475-489, (2011).

[3] Chakravarti et al. The organic grid: Self-organizing computation on a peer-to-peer network. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans (2005) vol. 35 (3) pp. 373-384

[4] Liabotis et al. Self-organising management of Grid environments. International Symposium on Telecommunications (2003)

[5] Spanish Academic and Research Network: **http://www.rediris.es/rediris/**

[6] Schwiegelshohn et al. Perspectives on grid computing. Future Generation Computer Systems (2010) vol. 26 (8) pp. 1104-1115.

[7] Muttoni et al. Optimal number of nodes for computation in grid environments. Proc. of the 12th Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP04) (2004) pp. 282289

[8] da Fontoura Costa et al. Complex grid computing. The European Physical Journal B (2005) vol. 44 (1) pp. 119-128

[9] Ishii et al. A Complex Network-Based Approach for Job Scheduling in Grid Environments. Lecture Notes in Computer Science (2007) vol. 4782 pp. 204

[10] Lazarevic. Probabilistic grid scheduling based on job statistics and monitoring information. Transfer thesis for the degree of Doctor of Philosophy. (2005)

[11] Clauset et al. Power-law distributions in empirical data. Arxiv preprint arXiv:0706.1062 (2007)

[12] Spanish e-Science National GRID Initiative Project **http://www.e-ciencia.es/**

[13] Barabasi. The physics of the Web. Physics World (2001) vol. 14 (7) pp. 3338