

# Distributed large-scale systems development: Exploring the collaborative development of the particle physics Grid

Avgousta Kyriakidou and Will Venters

London School of Economics and Political Sciences

Information Systems and Innovation Group, Department of Management

a.kyriakidou@lse.ac.uk

## Abstract

This paper examines the distributed development of Grid infrastructure in the particle physics setting. Specifically, the focus of concern is the collaborative practices employed by particle physicists in their attempt to develop a usable Grid with the aim to offer lessons to those involved in globally distributed systems development.

## 1 INTRODUCTION

Grid computing promises to distribute and share computing resources “on tap” and provide transparent communication and collaboration between virtual groups (Foster and Kesselman 2003). Yet developing and implementing such complex information infrastructures requires collaboration among a range of dispersed groups, and flexibility and adaptability to volatile requirements (Berman, Geoffrey et al. 2003). Here, we examine a case-study of Grid development within particle physics (PP), the LCG (Large hadron collider (LHC) Computing Grid), in an attempt to explore how such a large-scale distributed system is developed collaboratively in a global way in readiness for data from experiments at the recently launched LHC particle accelerator at CERN, Geneva. The PP community is well-known for the development of other cutting edge distributed systems (e.g. the web) and is itself highly distributed, so presenting a context where distinctive collaborative practices emerge.

Exploring this case we argue that the development of Grids poses new and underexplored opportunities for understanding collaborative global systems development. We particularly examine the collaborative practices of the PP community as it develops its Grid with the aim to offer answers into the wider context of distributed systems development and provide concrete practical recommendations for those considering the collaborative development of large-scale systems. Our theoretical framework is drawn from activity theory, and frames the LCG project as a complex activity system influenced by the context, the community’s rules, culture, history, past-experiences, shared-visions and collaborative practices (Nardi 1996). We understand the Grid’s development as a series of contradictions between the elements of this activity system, which are in a continuous process of getting resolved in order for the activity system to achieve stability and balance (Bertelsen 2003).

The following section reviews the literature on global software development (GSD). Section 3 and 4 present the theoretical framework and methodology. The case study then follows, after which analysis is presented. Finally, tentative conclusions for the systems development and Grid communities are provided.

## **2 LITERATURE REVIEW**

With the current trend of globalization and the problems of turbulent business environments (Herbsleb, Paulish et al. 2005), the IT industry has turned toward globally distributed software development in an attempt for the silver bullet of high-quality software delivered cheaply and quickly (Agerfalk and Fitzgerald 2006). Ongoing innovations in IT have made it possible to cooperate in a distributed fashion. Particular attention is however, being given to the opportunities and difficulties associated with sharing knowledge and transferring “best practices” within and across organizations (Orlikowski 2002). The global outsourcing literature suggests that distributed collaborative practices in systems development are important and timely (Yalaho 2006). Over the years the practices involved in GSD have undergone refinement and new more effective practices that focus on the collaborative factor have emerged such as agile practices (Nerur, Mahapatra et al. 2005).

GSD is considered to be the new paradigm in developing large-scale systems (Damian and Moitra 2006). However, there are still challenges involved in managing the development, such as communication issues and technical issues that need to be addressed (ibid). While literature stresses for practices and processes that are flexible and adaptable to the increasingly volatile requirements of the business environment (Highsmith and Cockburn 2001), Lee, Delone et al. (2006) argue that successful GSD requires not only flexibility but also rigor in order to cope with complex challenges and requirements of global projects. Furthermore, there is an ongoing debate which rejects the idea of agile practices as a silver bullet to the challenges of GSD (Parnas 2006).

With GSD the limitations of traditional systems development practices become even more obvious (Hanseth and Monteiro 1998). There is also a fundamental shift from the development of traditional IS to the development of global information infrastructures, such as the Grid (Foster and Kesselman 2003). Grid infrastructures should be seen and treated as large-scale and open as they demand collaborative development in a global environment (ibid), an environment characterized by high uncertainty and complexity and a continuous stream of improvisation, bricolage, drifting etc. Development often requires ad-hoc problem solving skills and creativity, skills which cannot easily be pre-planned (Ciborra 2002). GSD demands new, different development practices as the nature of the problem is now different. Long-cherished computer science principles and early systems development are therefore re-examined in the light of the new requirements.

## **3 THEORETICAL FRAMEWORK**

In contrast to the deterministic views inherent in much of the literature on Grids we employ Activity theory (AT) as an approach to help us look at how technology is collaboratively constructed to fulfil the objectives of a global community (Nardi 1996). AT has inspired a number of theoretical reflections on what information systems development is about (Kuutti 1991). AT provides a well developed framework for analyzing the complex dynamics of collaborative settings which typically involve interacting human and technical elements (Crawford and Hasan 2006). The concept of collectiveness and the notion of different actors sharing the same goals and constructing the same meanings are at the core of this theory (ibid), and are vital to our analysis of an “exceptional” community.

AT’s conceptual tools such as the activity triangle model (ATM) and the notion of expansive learning and contradictions (Engestrom 1987) allow us to create a rich understanding of the development activity and how this is influenced by the distributed context, the community’s rules, culture and collaborative practice (ATM), discover the breakdowns that may arise in the

development activity during a period of time (contradictions) and therefore capture the possibilities for expansive developmental transformations by questioning the existing practice and identifying the newly implemented models in practice (expansive learning).

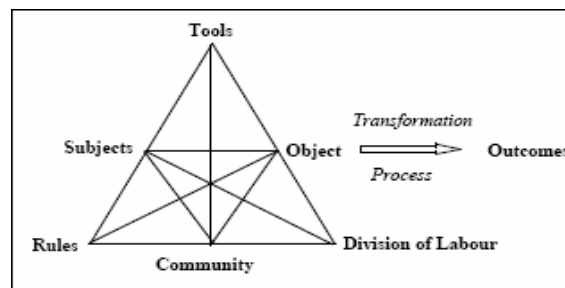


Figure 1 Activity triangle model (Engestrom 1987).

## 4 RESEARCH METHODOLOGY

### 4.1 Research context

Grid technology is claimed to be a fundamental step towards the realization of a common *service-oriented infrastructure* for on-demand, distributed, collaborative computing, based on open standards and open software (Foster and Kesselman 2003). Carr (2005) brashly suggests that the shift to Grid computing forms of technology will “overturn strategic assumptions, alter industrial economics, upset markets and pose daunting challenges for every user and vendor”. While these are obviously extremely bold predictions, and the term Grid remains ill-defined, Grids remain an important step towards global IT infrastructures. A Grid is just a large number of distributed processors and other computing devices linked through networks, and presented to the user as a single computer and without the need to address individual resources directly (unlike the web in which URLs address particular web-serving machines). Among many international Grid projects worldwide, PP stands out, because of their exceptional distributed collaboration (Chompalov, Genuth et al. 2002), their significant contribution to Grid’s development and the fit of their style of analysis to Grid’s capabilities.

### 4.2 Research design

LCG’s collaboration uniqueness, prevents comparative studies, but provides a revelatory case of distributed systems development practice. An interpretative case study is thus used to gain in-depth understanding of the dynamic, complex, loosely-coupled PP’s systems development activity and the collaborative construction of their shared practices. Research evidence was collected through over 70 semi-structured interviews with key members of LCG as well as observation of major meetings/workshops and three week-long trips to CERN. Reviewing of LCG’s documentation was also carried out. Interviews were audio-recorded, transcribed and coded with Atlas.ti, though this was used in a flexible means as a device to aid interpretation.

## 5 CASE STUDY: THE LCG PROJECT

On September 10<sup>th</sup> 2008 the Large Hadron Collider at CERN in Geneva began particle acceleration (though quickly stopped due to magnet failures). When fully operational this accelerator will produce six hundred million interactions per second and, after considerable filtering, 15 million Gigabytes of data annually, requiring a unique amount of processing power (100,000 CPUs) and storage space to allow the thousands of physicists globally to analyze them (Lloyd 2006). The proposed solution to this and other computational-intensive and data-centric problems is the Grid (ibid).

The LCG project’s mission is to build and maintain this Grid infrastructure for the entire high energy physics community that will use the LHC (ibid). Building the LCG is a highly distributed, complex and poorly defined systems development task. Cutting edge technology and tools are used, new standards reflecting security issues etc. are being negotiated and middleware (which is like the operating system of the Grid) together with other supporting software are being developed collaboratively by physicists around the world. Particle physicists (PPs) have a long tradition of such large-scale global collaborations and working on a distributed basis is just a part of their everyday routine (Knorr-Cetina 1999). Indeed, building this large-scale, by nature distributed Grid, demands global development and its development is organized into a number of projects, some of which extend beyond the physics community. Firstly, funding is so difficult to get and it is politics rather than technology which may inhibit the success of such Grid initiatives (Kyriakidou and Venters 2007) and secondly an enormous amount of manpower is needed for the different Grid elements to be developed. These dictate that Grid elements be globally distributed rather than collocated at CERN.

PPs’s collaborative work practices are not typical (Knorr-Cetina 1999; Shrum, Genuth et al. 2007) and have been described by Chompalov, Genuth et al. (2002) as “*exceptional*”. LCG’s constitution reflects these work practices and is thus based on a collaboration where decisions are made based on democratic and consensual basis with minimal levels of internal authority (Shrum, Genuth et al. 2007). The systems development activities undertaken by the LCG project varies. They include the development of middleware components, installation and maintenance of Grid hardware, development of physics applications for job submission to run on top of the Grid middleware, testing and certification of applications, user support. The Grid is already partially in use, and thus some physicists are already users who write software to undertake their analysis. Most meetings for coordinating and dealing with issues around the development are conducted virtually and everyday communication is achieved through email exchanges. Knowledge is located and socialized through these shared resources as well as through key individuals, who are considered experts and carry out such knowledge and expertise by attending different meetings and constantly changing job posts.

## 6 ANALYSIS

The scope of this paper is not to undertake a whole activity theory analysis and therefore provide in depth interpretations of the activity’s structure (activity-actions-operations). Rather, we aim to create a rich understanding of the LCG development activity and how this is influenced by the collaborative distributed context, the community’s norms and culture. We do this by constructing the developers’ activity system and so identifying some of the frictions and tensions through the AT concept of contradictions. Identifying an activity system (according to Engeström’s model) requires the identification of the following components:

<b><i>Activity of interest</i></b>	Distributed collaborative systems development
<b><i>Object</i></b>	Development of the Grid infrastructure for supporting the LHC
<b><i>Desired outcome</i></b>	LCG enabling analysis such that the LHC can achieve breakthroughs in physics
<b><i>Subject</i></b>	Particle physicists, computer scientists/software engineers
<b><i>Tools</i></b>	Systems development practices, programming languages, Draw on down-to-earth approaches embedded in PP tradition and history, Trial and error, reactionary and flexible approaches, improvisation, natural selection, hacked solutions
<b><i>Community</i></b>	Collaboration, belief in genius nature of work, high degree of competence, shared goals, trust, pragmatism etc.
<b><i>Rules</i></b>	Collaborative way of working, deliver on time within the budget
<b><i>Division of labour</i></b>	Fuzzy limit, everyone is doing what does best. People can have more than one jobs. Freedom, Limited lines of authority. No leader or manager, there are spokespersons

Table I: Developers’ activity system.

PPs are waiting for the LHC (**desired outcome**) to begin to fully operate. The objective behind the LCG's systems development activity was realized when particle physicists understood that in order for the LHC to be successful they needed a Grid to undertake analysis (**object**). This realization made PPs form global distributed virtual alliances with a number of actors such as funding bodies, universities and the industry. As one interviewee stated: *"What physicists want to do cannot be done by a small group, it needs a large collaboration"*. The development of this global infrastructure involves PPs with computing skills, as well as traditional computer scientists (**subjects**). As Grid technology is new and different and the complexity and scale of the project is such, PPs argue that they cannot take a plan-based approach to systems development. Their aim is to learn and move forward by trial and error and pragmatically and creatively react, drawing on the down-to earth and creative approaches (**tools**) embedded in their tradition and history.

PPs have themselves recognized their computing practices as amethodical, highly pragmatic and improvisational. The lack of formal processes in systems development is openly acknowledged, with most believing their existing work practices are effective given their primary purpose of building a working system in an extremely limited time-scale. PPs follow a bottom up and reactionary approach to development. However, such a distributed development environment requires flexibility and adaptability to changing requirements and external pressures and for this reason they make use of other more flexible practices. Developers have short-term goals and therefore have short cycles of iteration with continuous releases. Releases are usually tested and certified before they move into the pre-production Grid, where further robustness tests and feedback are gained. Developers also use prototypes for feedback, to improve functionalities and to gather requirements. Experience plays a crucial role as most of the developers are the physicists who themselves know what needs to be done. Concerning traditional systems development methodologies, the usage is little: As an interviewee claimed: *"I would say that we use methodologies but just up to the limit that it is appropriate for what we are trying to do...things change, how can you possibly do a formal software engineering approach to something if we are going to change it ?"*. A distinct feature of identifying and exploiting technical solutions in the project is the reliance on natural selection. Within PP the use of competing technological solutions is a traditional way of working – often as people simply try to solve their problems without consulting others (Pickering 1995). Physicists *"are powerful users, they do what they want. It's in their culture, it's their mentality.."*. Once these hacked solutions exist natural selection rather than politics or social power selects those to be carried forward: *"The cream comes to the top. Things that work win out and that's how we worked it"*.

Traweek (1988) described PPs as "promethean heroes of the search of the truth" and outlines the inherently collaborative nature of their community, with Knorr-Cetina (1999) similarly describing them as communitarian (**community**). Collaborative working has traces back in their history, their culture and the nature of their experiments seen as collaborations: *"We run big collaborations across every nation on earth and they always work"*. People take part in this global distributed collaboration and openly share their knowledge because they like feeling that they have contributed to the collective cause. They are driven by the shared goals and "sacred causes". As one interviewee put it *"This high level common goal [LHC] makes it actually easy for us to do this thing [to collaborate]"*.

PPs almost always state with certainty that the LCG will work because they are extremely clever and will make it work. A more significant source of confidence (one might argue arrogance) resides in the belief in the individual skills, competence, creativity and in the

context of collaboration. This high degree of competence, the shared goals and internal motivation as well as the collaborative working create a trustworthy environment which drives the community. Porra (1999) argues that the culture of collaboration or individualism within a colony is passed on from generation to generation as customs, norms, values, stories and behaviour - its rules of conduct. For PPs the collaborative way of working (**rules**) is inherent within individual physicists and this is the only, if we could say, “real rule” that guides them. Being so distributed it is crucial to build a strong sense of community and construct an identity for those involved in Grid’s development in order for the project to function collectively. Going to the pub or going together for lunch when co-located, for example, are one important aspect of this. Finally, there is no clear **division of labour** within the collaboration and individuals shift between jobs and have more than one job at the same time: “*We don’t regard our roles as having fixed boundaries... We tend to do things which we are better at regardless of whose role it might actually be.*”. Furthermore, they argue there is no strict hierarchy within the collaboration; what they argue they have is a collaboration with a spokesperson and volunteers rather than a company with a managing director and board. There is not a leader that directs what people are doing; rather people have freedom to improvise, use different techniques and have space for creativity and innovation.

## **7 STRATEGIES FOR SUCCESSFUL DISTRIBUTED DEVELOPMENT**

PPs acknowledge that the distributed development of the LCG has been a challenging learning journey. Drawing on AT we observe a number of contradictions emerging and being faced throughout the development process. These include i) tensions between particle physicists and computer scientists, ii) breakdowns between traditional technical models to development and what was actually going on etc. Their means of resolving these contradictions provide evidence of their “expansive learning” (in AT terms). Such learning drives the progress of the development, and not only gave rise to technological innovations to support distributed collaboration but also to new work practices which have enabled developers to better face the demands of such large-scale distributed development. It is in these practices that we find lessons for those engaged in the distributed development.

### **7.1 Clusters of competence.**

PPs are highly influenced by their past experiences and their practices are rooted in their history and culture. One of the lessons they learnt which led to changes in the way they work was the realization that fully distributed development is difficult because of problems of integrating work effectively: “*It is very difficult to have different teams that are distributed working on the same component because the elements they develop might not work together in the end*”. Full distribution of the work, originally believed to be the right way to approach the situation, presented a contradiction to what was actually going on in practice. Their answer to this contradiction was the idea of (in their own words) “Clusters of Competence” which enabled them to structure the development in different competent clusters: “*We are trying to go to a situation where one component is being developed by the same group of people who are all in one place*”. They have created different globally distributed patches of expertise, where experts are co-located, which are then all aligned into a network that facilitates and coordinates the work, the collaboration as well as the sharing of knowledge among the different clusters. However, although the clusters of competence were found to provide discipline in messy situations, for the development of some components of the Grid this was not possible and hence there are still people working in a truly distributed way. While virtual communication is important for standardization, PPs still encourage temporary co-location of developers: “*Developers from other countries come here for 3 months and this is much more efficient... There is still freedom but in a controlled way.*”

## 7.2 Balancing experimentation with discipline.

All PPs must write computer software in order to undertake their physics analysis since packaged applications for this task do not exist. However, they do not have formal training in software engineering: *“As a physicist you do not get much experience in writing software which stays up and is reliable”*. For them developing software is an experimental activity involving trial-and-error in a way similar to the way physics itself is undertaken. However, when asked about this unstructured “experimental” (and risky) way of working, they have all agreed that in such kinds of distributed development projects someone must combine this agility/flexibility together with limited structure/discipline: *“One thing the project learned is that you need management and clear short-term priorities, or else you drift”*. Because this project consists of both PPs and computer scientists, it is argued that some discipline in development activities is needed in order to balance the developers’ individual goals with the shared objective. However, it is agreed that it is also crucial to maintain this flexible and agile character in the way they work in order to quickly adapt and respond to changes.

## 7.3 A sense of belonging.

Interestingly, when asked what other communities could learn from the distributed development of their Grid, interviewees suggested that what makes their project progress is a combination of factors. As they have argued: *“We’ve many times seen the development of systems by isolated groups involving formal procedures. But these didn’t have good results”*. Therefore something more than co-location and formal procedures is needed in order for such kind of virtual projects to be successful: *“Social is the key really. It makes such a huge difference when people work together for the right reason. True quality comes from within”*. Indeed creating a strong sense of community with shared goals is crucial for their collaboration: *“Collaboration and building community is really important for distributed development. We work a lot using mailing lists; you can see the different attitude people have before and after they meet in person in those mailing lists”*. Shared goals provide motivation and an identity is constructed for those involved in the development of the Grid. The feeling of belonging to a group also balances competitive relationships: As they argue: *“Proper management of competition leads to successful outcomes. Without competition brilliant ideas are killed. That is why we work through competing solutions and the best one wins out”*.

## 8 CONCLUSIONS

This paper examined the collaborative practices employed in the distributed development of Grid infrastructure in a particle physics setting with the aim of offering answers that may be translated into the wider context of global virtual development. PPs appear both highly unusual and somewhat traditional in the way they work. Freedom, trust, consensus, charismatic leadership, shared goals and internal motivation are all distinct characteristics of the community and are seen to be its major driving forces. However, this is not to say that politics do not exist or that competition is minimized. Rather, healthy competition exists which helps bring competence in the community and blends expertise.

A number of strategies for distributed development were presented based on lessons PPs learnt throughout the Grid’s development. In summary the strategies identified were: 1) Structure the development effort in clusters of competence, 2) Encourage temporary co-location of developers, 3) Combine flexibility/agility with structure/discipline, 4) Create a sense of belonging and therefore construct identity for those involved in the development, 5) Create a trustworthy environment, 6) Have clear shared goals and rationale. These lessons resonate with many of the trends in management theory around effective distributed working, yet the fact that they emerge from a unique community founded not in existing bureaucratic

commercial organisations but in a communitarian science practices provides important evidence to this ongoing debate. Clearly, the unique and obscure nature of the community under study is also a limitation to the study. The practical recommendations provided here should not be seen or treated as prescriptive canonical rules but hopefully they will allow others involved in such distributed virtual development projects to reflect on their practice and their context in light of them.

**ACKNOWLEDGEMENT:** We would like to thank the LCG collaboration for providing generous access and assistance to our research. This research is undertaken as part of the Pegasus project (Particle-physics Engagement with the Grid: A Socio-technical Usability Study) funded by the UK EPSRC (Grant no EP/D049954/1). Further details available at: [www.pegasus.lse.ac.uk](http://www.pegasus.lse.ac.uk)

## 9 REFERENCES

- Agerfalk, J. P. and B. Fitzgerald (2006). "Flexible and distributed software processes: Old returns in new bowls?" Communications of the ACM **49**(10).
- Bertelsen, O. W. (2003). Contradictions as a tool in IT-design: Some notes. 8th European conference of computer supported cooperative work (ECSCW), Helsinki, Finland.
- Carr, N. (2005). "The End of Corporate Computing." MIT Sloan Management Review **46**(3): 67-73.
- Chompalov, I., J. Genuth, et al. (2002). "The organization of scientific collaborations." Research Policy **31**: 749-767.
- Ciborra, C. (2002). The Labyrinths of Information: Challenging the Wisdom of Systems. Oxford, UK, Oxford University Press.
- Cockburn, A. and J. Highsmith (2001). "Agile software development: The People factor." IEEE Computer
- Crawford, K. and E. Hasan (2006). "Demonstrations of the activity theory framework for research in information systems." Australian journal of information systems **13**(2): 49-68.
- Damian, D. and D. Moitra (2006). "Global software development: How far have we come?" IEEE Software.
- Engestrom, Y. (1987). Learning by expanding: An activity-theoretical approach to developmental research. Helsinki, Orienta-Konsultit.
- Foster, I. and C. Kesselman (2003). The Grid in a Nutshell. In: Grid Resource Management – State of the Art and Future Trends. J. Nabrzyski, J. Schopf and J. Weglarz, Kluwer Academic Publishers.
- Hanseth, O. and E. Monteiro. (1998). "Understanding Information Infrastructure." <http://heim.ifi.uio.no/~oleha/Publications/bok.html> Retrieved 15th of June, 2006.
- Herbsleb, D. J., J. D. Paulish, et al. (2005). Global software development at Siemens: experience from nine projects. International conference on Software engineering (ICSE'05). USA, ACM.
- Highsmith, J. and A. Cockburn (2001). "Agile software development: The business of innovation." IEEE Computer Society.
- Knorr-Cetina, K. (1999). Epistemic Cultures: How the sciences make knowledge. Cambridge, MA, Harvard University Press.
- Kuutti, K. (1991). Activity theory and Its Applications to Information Systems Research and Development. Information Systems Research: Contemporary Approaches & Emergent Traditions. H. E. Nissen, H. K. Klein and R. Hirschheim, Amsterdam: North-Holland: 529-550.
- Kyriakidou, A and W. Venters (2007). "The multi-disciplinary development of collaborative Grids: The social shaping of a Grid for healthcare", 15th European Conference on Information Systems (nominated for best paper), St Gallen, Switzerland
- Lee, G., W. Delone, et al. (2006). "Ambidextrous coping strategies in globally distributed software development projects." Communications of the ACM **49**(10).
- Lloyd, S. (2006). From Web to the Grid. PPARC and Parliament and "House" magazine.
- Nardi, B. (1996). Context and Consciousness: Activity theory and human computer interaction. Cambridge, MIT Publisher.

- Nerur, S., R. Mahapatra, et al. (2005). "Challenges of Migrating to Agile Methodologies." Communications of the ACM **48**(5).
- Orlikowski, W. J. (2002). "Knowing in practice: Enacting collective capability in distributed organizing." Organization Science **13**(3): 249-273.
- Parnas, D. (2006). "Agile methods and global software development (GSD): The wrong solution to an old but real problem." Communications of the ACM **49**(10).
- Porra, J. (1999). "Colonial systems." Information Systems Research **10**(1).
- Shrum, W., J. Genuth, et al. (2007). Structures of Scientific Collaboration. Cambridge, MIT Press.
- Traweek, S. (1988). Beamtimes and lifetimes: The world of high energy physics. Cambridge MA, Harvard University Press.
- Yalaho, A. (2006). A conceptual model of ICT-supported unified process of international outsourcing of software production. 10th IEEE International Enterprise Distributed Object Computing Conference Workshops (EDOCW'06), IEEE.