

Meeting the Design Challenges of Nano-CMOS Electronics:

An Introduction to an Upcoming EPSRC Pilot Project

R. Sinnott¹, A. Asenov², D. Berry³, D. Cumming², S. Furber⁴, C. Millar², A. Murray⁵,
S. Pickles⁶, S. Roy², A. Tyrrell⁷, M. Zwolinski⁸

¹National e-Science Centre, University of Glasgow

²Department of Electronics and Electrical Engineering, University of Glasgow

³National e-Science Centre, University of Edinburgh

⁴Advanced Processor Technologies Group, University of Manchester

⁵Mixed Mode Design Group, University of Edinburgh

⁶e-Science North West, University of Manchester

⁷Intelligent Systems Group, University of York

⁸Electronic Systems Design Group, University of Southampton

ros@dcs.gla.ac.uk

Abstract

The years of 'happy scaling' are over and the fundamental challenges that the semiconductor industry faces, at both technology and device level, will impinge deeply upon the design of future integrated circuits and systems. This paper provides an introduction to these challenges and gives an overview of the Grid infrastructure that will be developed as part of a recently funded EPSRC pilot project to address them, and we hope, which will revolutionise the electronics design industry.

1. Introduction

Progressive scaling of complementary metal oxide semiconductor (CMOS) transistors, as tracked by the International Technology Roadmap for Semiconductors (ITRS) [1] and captured in Moore's law, has driven the phenomenal success of the semiconductor industry, delivering larger, faster, cheaper circuits. Silicon technology has now entered the nano-CMOS era with 40 nm MOSFETs in mass production at the current 90 nm ITRS technology node [2] and sub-10 nm transistors expected at the 22 nm technology node, scheduled for production in 2018. 4 nm transistors have already been demonstrated experimentally [3], highlighting silicon's potential for scaling beyond the end of the current ITRS. However, it is widely recognised that variability in device characteristics and the need to introduce novel device architectures represent major challenges to scaling and integration for present and next generation nano-CMOS transistors and circuits. This will in turn demand revolutionary changes in the way in which future integrated circuits and systems are designed. To tackle this problem, strong links must be established between circuit design, system design and fundamental device technology to allow circuits and systems to accommodate the individual behaviour of every transistor on a chip.

Design paradigms must change to accommodate this increasing variability. Adjusting for new device architectures and device variability will add significant complexity to the design process, requiring orchestration of a broad spectrum of design tools by geographically distributed teams of device experts, circuit and system designers. This can only be achieved by embedding e-Science technology and know-how across the whole nano-CMOS electronics design process and revolutionising the way in which these disparate groups currently work. The recently funded "Meeting the Design Challenges of Nano-CMOS Electronics" EPSRC pilot project is looking directly at building a Grid infrastructure that will meet the challenges raised by the scaling problems across the whole of the electronics industry. This 4-year project is expected to start in October 2006 hence this paper is focused upon the domain requirements and scientific challenges that will shape the Grid infrastructure. We also present initial ideas in the design and implementation of the Grid infrastructure that will address the specific challenges of this domain.

The rest of the paper is structured as follows. Section 2 focuses on the scientific demands of the nano-CMOS electronics area and the problems arising from decreasing transistor scalability. Section 3 gives an overview of the demands that this domain places on the Grid infrastructure to be developed. Section 4 focuses on initial ideas on the design and development of this infrastructure, and we conclude with a summary of our plans for the future.

2. Scientific Challenges

The rapid increase in intrinsic parameter fluctuations represents the most serious challenge facing the electronics industry today. These fluctuations stem from the fundamental discreteness of charge and matter. They are fundamental, truly stochastic and cannot be eliminated by tighter process control. The major sources of intrinsic parameter fluctuations include random discrete dopants (Fig. 1 and Fig. 2), line edge roughness and oxide thickness fluctuations [4].

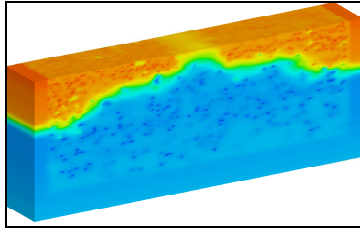


Fig. 1 Random discrete dopants in a 35 nm MOSFET from the present 90 nm technology node

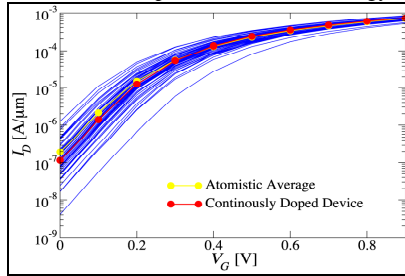


Fig. 2 Corresponding variations in the current-voltage characteristics of 200 transistors with different dopant distributions

While intrinsic parameter fluctuations and resultant device mismatch have hitherto affected only analogue design, they now challenge the power consumption, yield and reliability of digital circuits [5]. One of the first digital “casualties” is SRAM, which occupies significant real estate in current System On Chip (SoC) devices [6]. Fig. 3 illustrates the random dopant induced distribution of static noise margin in an ensemble of SRAM cells of various cell ratios at the transition between the 90 nm and 65 nm technology nodes.

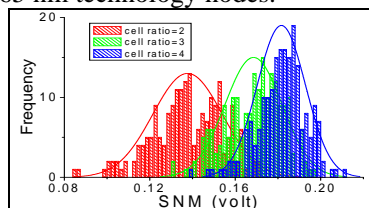


Fig. 3 Corresponding distribution of the static noise margins in 6T SRAM cells

Only a large cell ratio can produce acceptable yield in the presence of fluctuations, increasing cell area and reducing the benefits of scaling. Thus, variability already causes significant circuit and system challenges at a time when design margins are shrinking, owing to lowered VDD and increased transistor count. Exponentially increasing design difficulties require novel statistical design solutions. This is exacerbated by the enormous logistic, computational and data management needs of statistical design techniques which thus represents a prime candidate for exploitation of Grid technology.

In the years of ‘happy scaling’, circuit and system design used conventional bulk MOSFETs that behaved remarkably similarly over many technology node generations. Variability was associated with fabrication processes and equipment-related non-uniformities. Differences in, for example, implantation dosage and lithographical alignment were responsible for wafer-to-wafer parameter variations, and on-wafer non-uniformities were responsible for on-wafer variations. Simple workstation based ‘corner’ analysis was able to assess the impact of variations in the design process. As a result, compact models extracted from measured device characteristics supported by simple rule-based tools allowed a high level of abstraction, distancing circuit and systems designers from device design and technology.

New types of device parameter variations, related to the introduction of sub-wavelength lithography and process induced strain, emerged in the transition from the 130 to 90 nm technology nodes. These now play an increasingly important role. Optical proximity correction (OPC) and phase shift masks result in variations in the shape of transistors with otherwise identically drawn gate layouts, depending upon the surrounding cell topology. These dimensional variations are commensurable with the gate length and can result in significant changes in transistor characteristics. Compressive and tensile strain, induced typically by SiGe source/drain regions and Si_3N_4 blankets respectively, were introduced at the 90 nm

node to improve p- and n-MOSFET mobility and performance [2]. The strain distribution and device performance are determined by not only gate topology, but also by the gate-to-gate spacing, the width of the source/drain regions, the position and shape of the contact windows and the distance to the shallow trench isolation. OPC and strain-induced variations at and beyond the 65 nm node mean that standard design rules and conventional physical verification may not be sufficient to ensure yield without an unacceptable degradation in cell density. At the 45 nm technology node, hundreds of pages of design rules are expected to replace the traditional single page of rules, in order to maintain yield. Variations must be considered early in the design flow. Furthermore, the strong link between circuit and device design and underpinning technology design that was broken, for good reasons, in the early days of VLSI design must be re-established.

It is expected that there will be no single replacement for conventional MOSFETs and that disparate device architectures will coexist and compete. All new device architectures require a more-or-less new design approach, altering device and circuit layout and the electrical behaviour of each generation of nano-CMOS devices. This adds to the design challenges associated with increasing device and circuit variability.

Grid technologies when correctly applied across the nano-CMOS electronics design space can address these challenges. These infrastructures should allow designers to choose the most appropriate technologies for each design project, with the resources needed to deliver optimal, competitive design solutions. Importantly in this domain (which is one of the distinguishing features from other domains) is the importance of intellectual property (IP). IP for designs, data and processes is fundamental to this domain and SMEs and collaborators must be assured that the security infrastructure supporting the new design processes fully protects IP integrity.

3. Grid Challenges

Whilst there are numerous areas where we expect to extend state of the art in Grid

know-how, our fundamental goal is to facilitate scientific research: ideally to revolutionise the electronics design industry.

We have identified several areas that characterise capabilities of the infrastructure needed to support the scientific challenges of the nano-CMOS domain. For each of these areas we envisage developing a family of components comprising frameworks that can be applied by the scientists for their daily research, incorporating all aspects of the designs of circuits and systems incorporating the decreasing scaling and expanding design capabilities that face the electronics industry described previously.

Specifically in exploratory domain discussions on capabilities needed by the electronics design protagonists we have identified the following four key areas that are crucial to the success of the Grid infrastructure: workflows, distributed security, data management, and resource management.

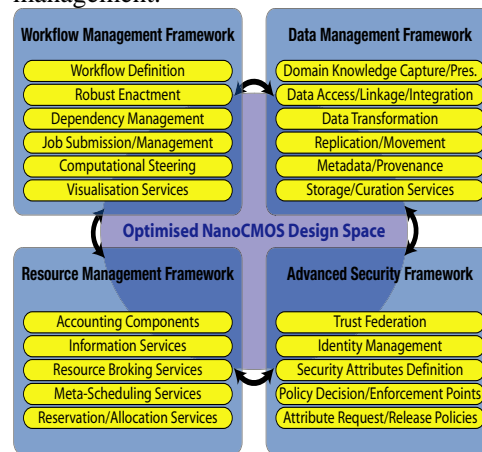


Figure 4: Framework and Components of the Grid Infrastructure

We consider these in turn and why they are important, and provide an initial overview of our intentions in delivering these components.

3.1 Workflow Management Framework

The definition and enactment of workflows suitable for the nano-CMOS electronics domain will form the cornerstone of our work. This will require the wrapping of existing simulation software as Grid services by the application design groups, aided by the e-

Science partners. Figure 5 indicates our initial expectations on the kinds of workflow components that need to be supported.

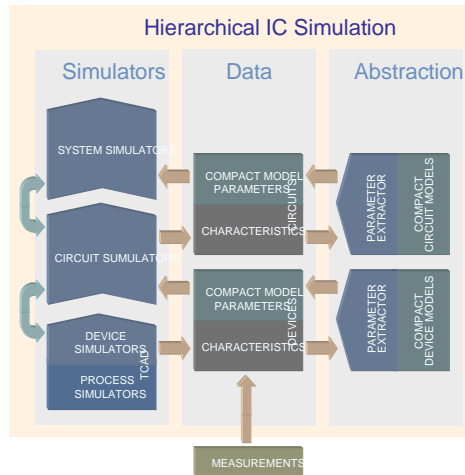


Fig. 5 Hierarchical simulation methodology needed to capture the impact of variability on design.

We recognise already an important issue here will be to describe the services and data sets with sufficient information (meta-data) to allow for their subsequent discovery, access and usage in the design workflows. Given the multitude of atomistic device simulation software variations, where each simulation is compiled to explore slightly different parameter sets, ensuring that the similarities and distinctions between these simulations and their input/output data sets are captured is fundamental to the workflow definition and their enactment. Where possible, the design groups will automatically capture this meta-data, building on existing tools and expertise from the e-Science partners as described below in section 3.3.

We plan to allow design groups to work together and develop libraries of workflows that allow other scientists to run, and subsequently manage, multiple, concurrently executing simulations. The scientists will be able to browse and search for workflows, designs and data sets relevant to their work, subject to their security attributes.

We plan to allow for broad parameter sweep simulations which, through user interaction, can be refined to fine-grained

parameter explorations when a situation of specific scientific interest arises. It will also allow for the efficient overall usage of the Grid infrastructure resources, since uninteresting simulations can be stopped or steered to more interesting regions of parameter space. Services supporting visualisation of scientific data will be integrated into these workflows. Interactive steering in a workflow context will be a novel application of RealityGrid (www.realitygrid.org.uk) and myGrid (www.mygrid.ork.uk) software, and their integration will give rise to new requirements on both. As the RealityGrid software is maintained by project partners, we can deal with requirements on it ourselves. However, future development of myGrid software is the responsibility of OMII-UK (www.omii.ac.uk), and requirements on the myGrid provenance service and Taverna workflow enactment engine or their successors must be considered in a broader context. In order to manage this external dependency, we will establish an ongoing dialog with OMII-UK at the project start, negotiate mechanisms for feeding our requirements into the OMII-UK roadmap, and if necessary, contribute effort to their realisation and verification.

3.2 Resource Management Framework

The optimised use of the compute resources requires up to date information on the status of the Grid infrastructure. To support this, the e-Science partners will deploy Grid services which will capture near real time data on the status of the Grid infrastructure and the associated Grid services deployed. We will adapt and extend existing services, e.g. from the NeSC BRIDGES project (www.nesc.ac.uk/hub/projects/bridges) and the Globus Alliance web service based monitoring and discovery system (MDS) [7], while taking into account developments on the National Grid Service (NGS www.ngs.ac.uk). These services will include capabilities for publishing and subscribing to information service data sets, for filtering of the associated data with these services and for storing and archiving the data associated with these services. Aggregation of such

information will be supported and incorporated within the workflow management framework to influence real time workflow enactment. The definition of appropriate schemas will be fundamental to this work and we will build on the OGSA working group's current comparison of GLUE and DMTF CIM, where the e-Science partners are intimately involved [8].

Understanding where a given simulation should be executed can be an involved process involving numerous metrics, e.g. the status of the Grid infrastructure at that time, the chip architecture that the code has been compiled for, the location of specific data sets, the expected duration of the job, the authorisation credentials of the end user wishing to run the workflow etc. To support the scientific needs of the project, we will survey on-going work in this area, and if necessary develop and deploy our own meta-scheduling and planning services building on the basic meta-scheduler currently supported in the NeSC BRIDGES project.

Metrics on data movement, as well as existing and predicted resource usage will be explored as part of this work. This will include the exploration of different economic models, such as maximal job throughput, minimal job cost in the presence of separately owned and managed Grid resources, each with their own specific accounting and charging policies.

Given that the fEC model of research funding requires monies to be set aside for time spent on major HPC clusters, we propose to explore existing state of the art in Grid based accounting services. One of these which we will explore and potentially enhance is the Resource Usage Service [9] which is currently being considered for deployment across the NGS, and will be considered for the (www.scotgrid.ac.uk) and the Northwest Grid. We plan here to draw on expertise and software from the Market for Computational Services project (see <http://www.lesc.ic.ac.uk/markets/>). In the latter part of the project, we also plan to explore advanced resource broking, reservation and allocation. We would

expect to learn from and feed into the GRAAP-WG at GGF and on-going efforts in this area, such as the WS-Agreement standard specification [10]. The case for advanced reservation and allocation will be tempered by the practical impact on reduction of overall utilisation of the compute resources and the associated cost impact this will incur. A better understanding of these issues is crucial in the fEC era.

3.3 Data Management Framework

The data sets that are generated by device modellers and circuit/systems designers are significant, both in size and in number as indicated in table 1.

| Task | Accumulated data project lifetime |
|--|-----------------------------------|
| SPICE cell & cct char. | 10GB |
| T-level sim. (nanosim) | 5-10TB |
| Gate-level sim. (Verilog) | 1TB |
| Behavioural sim. (Sys C) | 100GB |
| Extraction | 20GB |
| 3D TCAD sim. | 1TB |
| 3D 'atomistic' sim. | 5-10TB |
| Compact models | 30GB |
| Circuit level fault sim. | 100GB |
| Behavioural sim. | 10GB |
| Evolutionary systems | 30GB |
| Fragment sims. | 10 TB |
| Cells sims. | 5 TB |
| Extraction to STA | 100 GB |
| Sim. mixed-mode circuits with noise and variation. | 5TB |

Table 1: Summary of Expected Data Set Accumulation

The tight linkage and integration of these data sets and models is paramount. We plan to use and extend the OGSA-DAI system (www.ogsadai.org.uk) and the OGSA Data Architecture efforts to meet this need, including the current work in OGSA-DAI to manage files as well as databases. Particular areas where we will focus will include: (i) the integration of OGSA-DAI with workflow systems including data transformation capabilities, (ii) the development of appropriate meta-data schemas specific to the electronics domain, (iii) attaching security restrictions to data as it is moved (rather than to the sources and sinks themselves), (iv) comparison of remote data access with

pre-staging or the movement of application code to the data, (v) the efficient specification of data transfer for a variety of endpoints (e.g. files, query results, in-memory data sets), (vi) the integration of OGSA-DAI and provenance systems.

Annotating data from simulations with appropriate meta-data will allow for future tracking and longer-term curation and will form a fundamental part of the data management framework. Building on the close synergy of NeSC with the National Digital Curation Centre (DCC www.dcc.ac.uk) we will exploit direct expertise in how best to capture and subsequently exploit such information.

Some of the data sets, software and processes will be of commercially sensitive nature, and access to them must be restricted to suitably authorized individuals; when such data is transported or replicated, it must be done so securely, and in extreme cases, the data may not be replicated outside the domain in which it originated. Our data management framework will be closely coupled with the security framework to ensure IP is handled appropriately.

3.4 Advanced Security Framework

Novel device designs and their potential impact on integrated systems give rise to highly sensitive, commercial exploitation possibilities. Without a robust, reliable and simple Grid security infrastructure (from the end user perspective) incorporating very fine grained security capabilities, the electronics design community, from SMEs to major corporations involved in the electronics industry, will not involve themselves. The widespread acceptance and uptake of Grid technology can only be achieved, if it can be demonstrated that security mechanisms needed to support Grid based collaborations are at least as strong as local security mechanisms.

Drawing on the e-Health projects at NeSC, we will show how data security, confidentiality and rights management can be supported by the Grid infrastructure to protect commercial IP interests.

The predominant, current, method by which security is addressed in the Grid

community is through Public Key Infrastructures (PKI) to support authentication. This addresses user identity issues, but for the fine-grained control required over users' privileges on remote resources, we require advanced authorisation services. The project partners bring a wealth of experience in the practical establishment and management of advanced privilege management infrastructures using the latest advances in solutions such as PERMIS (www.permis.org). Additionally, the UK academic community is in the process of deploying the Internet2 Shibboleth technologies (shibboleth.internet2.edu) to support local (existing) methods of authentication for remote login to resources. Within this proposal we will explore the use of Shibboleth technologies to simplify the overall end user experience of access to, and usage of, Grid resources, drawing upon numerous other projects. We will identify and define the security attributes required in the electronics design domain. A direct application of this security infrastructure will be to restrict access to, and usage of, data sets and software which have IP restrictions. These are novel challenges and remain open issues to be solved within the Grid community.

3.5 Grid Summary

The seamless orchestration of the four frameworks and their components will create a *virtual nano-CMOS design foundry* where the behaviour of advanced systems and circuits can be predicted based upon, and feeding back into, device models and processes. The Grid infrastructure will allow the exploration of interesting challenges arising from situations where designs have been identified as invalid, erroneous or superseded. Tracking data sets and relations between the design space and models, whilst keeping the data design space as accurate as possible, is novel research in itself. Further challenges will be encountered in ensuring that IP issues and associated policies are demonstrably enforced by the infrastructure. The results of this work will directly impact upon future Grid efforts in the standardisation

and implementation areas. We expect to directly input the security solutions incorporating Shibboleth and advanced authorisation into OMII-UK version 5 releases (currently scheduled for 2007 in the draft roadmap) and provide a rigorous assessment and feedback on their workflow and enactment engine and their enhancements.

4. Initial Design Scenarios

To understand how these frameworks and components will be applied to support the nanoCMOS Grid infrastructure we outline one of the key scenarios that we intend to support. The requirements capture, design and prototyping phases that run through the lifetime of the project will refine this scenario and produce numerous other scenarios.

We consider a scenario where a scientist wishes to run an atomistic device modelling simulation based on the commercial Taurus software which requires a software licence to generate statistically significant sets of I/V curves. These I/V curves will then be fed into the Aurora software to generate compact models which will subsequently be fed into a circuit analysis software package such as SPICE. At each of these stages the scientists will be able to see the results of the software production runs and influence their behaviour. Note that Taurus, Aurora and SPICE are representative examples only and a far richer set of simulation software will be supported.

Step 1: A Glasgow researcher attempts to log in to the project portal and is automatically redirected via Shibboleth to a WAYF service (we will utilise the SDSS Federation (www.sdss.ac.uk)) and authenticates themselves at their home institution with their home username/password (denoted here by LDAP server – as used at Glasgow).

Step 2: The security attributes agreed within the nanoCMOS federation are returned and used to configure the portal, i.e. they see the services (portlets which will access the Grid services) they are authorised for. These security attributes will include licenses for software they possess at their home institution and their role in the federation amongst others.

Step 3: A client portlet for the Taurus software is selected by the scientist.

Step 4: The scientist then explores and accesses the virtualised data design space for input data sets to the Taurus software production run. This might consist of experimental data, first principle simulation data or data stemming from circuit or system inputs. Once again, what the scientist can see is determined by their security privileges (to protect IP). The meta-data describing the characteristics of the data such as the confidence rating (whether it has been validated or been superseded), who created it, when it was created, what software (or software pipelines) and which versions of the software were used to create the data, whether there are any IP issues associated with the data will all be crucial here.

Step 5: Once the user has selected the appropriate data sets needed for generation of the appropriate I/V curves, the meta-scheduler/planner is contacted to submit the job. Where the jobs are submitted will depend on which sites have access to an appropriate license for the Taurus software as well as the existing information on the state of the Grid at that time.

Step 6: Once the meta-scheduler/planner submits the jobs to the Grid resources and the portlet is updated with real time information on the status of the jobs that have been submitted (whether they are queued, running, completed). The actual job submission might be involved here, for example when the input files are very large and require to be partitioned. We will draw on the JSDL standards work here, e.g. through the OMII GridSAM software.

Step 7: On completion (or during the running of the Taurus simulations), the resultant I/V data sets are either stored along with all appropriate meta-data (not shown here) or fed directly (potentially via appropriate data format transformation services not shown here) into the Aurora software which in turn will decide where best to run the Aurora simulation jobs using the meta-scheduler/planner and available run time information.

Step 8: The Aurora client portlet will allow for monitoring of the resultant compact models and allow these to be fed

into the SPICE models (also started and run using the meta-scheduler/planner).

This orchestration of the different simulations and how they can feed directly into one another typifies the capabilities of the Grid infrastructure we will support and indicates how we will support a virtual nanoCMOS electronics design space. The e-Science partners in the project will initially design families of such workflows which the scientists can parameterise and use for their daily research, however as the scientists become more experienced in the Grid technologies, they will design and make available their own workflows for others to use.

Interesting challenges that arise from this domain that the Grid infrastructure will allow to explore will be when designs have been recognised to be invalid, erroneous or superseded. Tracking data sets and relations between the design space and models to keep the data design space as accurate as possible is novel. Further key challenges will be to ensure that IP issues and associated policies are demonstrably enforced by the infrastructure. Drawing on the e-Health projects at NeSC, we will show how data security, confidentiality and rights management can be supported by the Grid infrastructure to protect commercial IP interests.

5. Conclusions

The electronics design industry is facing numerous challenges caused by the decreasing scale of transistors and the increase in device design flexibility. The challenges whilst great are not insurmountable. We believe that through a Grid infrastructure and associated know-how, a radical change in the design practices of the electronic design teams can be achieved.

To address these challenges, the scientists cannot work in isolation, but must address the issues in circuit and systems design in conjunction with the atomic level aspects of nano-scale transistors. We also emphasise that the success of this project will not be based upon development of a Grid infrastructure alone. It requires the electronics design community “as a whole” to engage. This

can only be achieved if they are integrally involved in the design of this infrastructure. To achieve this we plan to educate the electronics design teams to an extent whereby they can Grid enable their own software, design their own workflows, annotate their own data sets etc. It is only by the successful adoption of these kinds of practices that the infrastructure will “revolutionise the electronics design industry” as we hope.

6. References

1. International Technology Roadmap for Semiconductors Sematech <http://public.itrs.net>
2. R. Khumakear et al., “An enhanced 90nm High Performance technology with Strong Performance Improvement from Stress and Mobility Increase through Simple Process Changes” 2004 Symposium on VLSI Technology, Digest of Technical Papers, pp 162-163, 2004
3. H. Wakabayashi, “Sub 10-nm Planar-Bulk-CMOS Device using Lateral Junction Control”, IEDM Tech. Digest, pp. 989-991, 2003.
4. A. Asenov, A. R. Brown, J. H. Davies, S. Kaya and G. Slavcheva, “Simulation of Intrinsic Parameter Fluctuations in Decanometre and Nanometre scale MOSFETs”, IEEE Trans. on Electron Devices, Vol.50, No.9, pp.1837-1852, 2003.
5. P.A. Stolk, H.P. Tuinhout, R. Duffy, et al., “CMOS Device Optimisation for Mixed-Signal Technologies”, IEDM Tech Digest, pp.215-218, 2001
6. B. Cheng, S. Roy, G. Roy, F. Adamu-Lema and A. Asenov, “Impact of Intrinsic Parameter Fluctuations in Decanano MOSFETs on Yield and Functionality of SRAM Cells”, Solid-State Electronics, Vol. 49, pp.740-746, 2005.
7. Globus Toolkit Monitoring and Discovery System, www.globus.org/toolkit/mds
8. Open Grid Service Architecture Common Information Model, www.ggf.org/cim
9. Resource Usage Service Working Group, www.doc.ic.ac.uk/~sjn5/GGF/rus-wg.html
10. A. Andrieux, K. Czajkowski, A. Dan, K. Keahey, H. Ludwig, J. Pruyne, J. Rofrano, S. Tuecke, and M. Xu. Web services agreement specification WS-Agreement (draft), 2004.