Flexible Software Technology for Scalable Cloud Computing

Carlos A. Varela
Worldwide Computing Laboratory, Department of Computer Science,
Rensselaer Polytechnic Institute, Troy, NY, U.S.A.
cvarela@cs.rpi.edu
http://wcl.cs.rpi.edu/

Topics

Programming Models for the Cloud
Cloud Self-Monitoring and Autonomic Control

Current Research Activities

Our research efforts include the design and development of advanced programming models, languages, and systems for adaptable fault-tolerant provably-correct distributed software. During the last ten years, we have worked on an actor-oriented programming language, called *SALSA*, which enables dynamic application reconfiguration [Varela and Agha, OOPSLA 2001] and can therefore be used to develop systems that scale well over cloud computing environments. In collaboration with IBM Research, we have also developed a fault-tolerant programming model, called *transactors (or τ-calculus,)* that enables reasoning about failures in distributed computations [Field and Varela, POPL 2005], and could be used as the basis for higher-level programming languages applicable to cloud computing. We are currently exploring further actor-based programming abstractions, including actor groups, to be used as a unit of mobility and failure.
In terms of distributed run-time systems, we have worked on a middleware layer, called the *Internet Operating System* (or *IOS*), to facilitate the autonomous dynamic reconfiguration of actors or processes, in order to improve performance and tolerate soft failures [El Maghraoui, Desell, Szymanski, and Varela, IJHPCA 2006]. We are currently exploring virtual machine migration and malleability [Wang, Huang, and Varela, Grid 2010] as reconfiguration operators for the middleware. We plan to extend IOS into a *Cloud Operating System* (COS), that can help cloud computing users save money and improve performance, while also enabling cloud computing providers to reduce energy usage and costs without sacrificing quality of service. Please see associated publications at: http://wcl.cs.rpi.edu/bib

Background and Experience

Dr. Varela received an NSF CAREER award on middleware and programming technology for grid computing in 2005. This award has led to three doctoral dissertations, and numerous publications including three best paper recognitions. Dr. Varela is currently Program Chair for the IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid 2011), and has been Associate Editor for ACM Computing Surveys since 2007.

Vision

New and emerging computing environments, from multi-core architectures to cloud computing, require software development techniques that enable fast human-friendly development of correct, efficient, and flexible concurrent software.

Flexibility of software is critical to meet the demands of cloud computing given the users' expectation for on-demand scalability. For example, software flexibility must enable to execute a large-scale scientific simulation partly on a private grid computing resource, and partly on a public cloud or volunteer computing grid, according to specified declarative user policies, which may favor cost, time, or energy.

Software flexibility must be supported at multiple levels of the software life cycle: at the programming layer, at the compilation layer, and at run-time through middleware for adaptive distributed systems execution.

Programming support for adaptive scalable cloud computing applications requires new research in: (1) high-level programming abstractions, (2) corresponding formal models with precise semantics that enable reasoning about system properties, (3) modal logics to easily specify and prove properties about concurrent systems as they relate to time, space, and knowledge, and (4) associated programming languages and tools.

Compilation support for adaptive scalable cloud computing applications requires new research in: (1) intermediate abstract code representations that enable targeting a diverse set of architectures including multi-cores, graphical processing units, and resource-constrained devices, (2) static typing and static code analysis techniques that guarantee certain errors will not occur at run-time even in the presence of dynamic software adaptation, and (3) code partitioning analyses that can help the run-time layers determine best resource allocation policies given a heterogeneous, dynamic, and potentially error-prone run-time environment.

Run-time support for adaptive scalable cloud computing applications requires new research in: (1) decentralized middleware, such as distributed operating systems, that can profile dynamic computing resources, that can profile distributed cloud applications progress, and that can decide on when to apply software reconfiguration to optimize time-to-solution, cost, energy, or a combination thereof, according to high-level user and provider policies, (2) multi-scale software reconfiguration techniques, including migration, replication, and malleability of software components at multiple scales including at the process level, at the virtual machine level and at the virtual network level, and (3) fundamental principles on autonomic computing, that is, applying control theoretic approaches to distributed pervasive systems adaptation, and reasoning about performance, stability, and optimality of dynamic reconfiguration techniques.