

## Research Statement

My research applies overlay networking and virtual machine technology to create easily deployable and self-managing systems for Wide-area distributed computing.

### 1. Background

Sharing of computing and storage resources between different institutions and individuals connected over the Internet is seen as a solution to meet the ever-increasing computation and storage demands of modern applications from different domains, such as high-energy physics, medical imaging and business data analysis, among others.

Middleware solutions have been proposed to facilitate resource sharing in a way that not only respects the policies defined by the resource owners, but also provides maximum flexibility to consumers. Systems have also been conceived and implemented to harness idle cycles from desktops of users connected to the Internet. Common to these efforts is the vision of providing computing as a utility that can be delivered by a pool of distributed resources in a seamless manner. The terms Grid and Utility Computing are used to refer to such systems for wide-area distributed computing.

### 2. Challenges

Several factors hinder deployment of unmodified applications over Wide-area. Firstly, heterogeneous resource configurations (including O/S kernels, libraries) may not be compatible with application or middleware requirements. Secondly, the increasing use of Network Address Translators (NATs) and Firewalls at Internet sites hinder legitimate access to shared resources. Finally, the inability to express and enforce sharing policies, and the lack of isolation provided by operating systems, greatly limit the range of applications and users that can benefit from resource sharing. Classic virtual machines (VMs) provide an excellent solution to overcome the compute resource heterogeneity, to enforce sharing policies and provide isolation from malicious applications. In [7], we have presented a Grid service that allows provisioning of such application-centric VM environments in less than a minute, on x86-based hosts with a suitable Virtual Machine Monitor (VMM). However, providing network access to resources in an unobtrusive manner, when majority of them are behind NATs and Firewalls, is still a challenge.

In addition, the scale and decentralized ownership of resources make management of Grid infrastructures extremely difficult. Recently, there is a trend towards self-managing (or self-\*) or autonomous systems that can function with the desired QoS guarantees in highly dynamic environments, even with minimal or almost no manual intervention. To realize the vision of providing computing as a utility, Grid infrastructures have to be made self-managing.

### 3. Current Research – Network virtualization

My current research addresses the problem of providing bi-directional network connectivity among wide-area resources behind NATs and Firewalls. Together with Dr. Renato Figueiredo and Dr. Oscar Boykin, I have developed a self-managing networking infrastructure -- IPOP [2] -- that aggregates wide-area hosts into a private network with decoupled address space management. The virtual network is functionally equivalent to a Local-area network (LAN) environment where a wealth of existing, unmodified IP-based applications can be deployed. The IPOP virtual network tunnels the traffic generated by applications over a P2P-based overlay provided by the Brunet P2P library, which handles NAT/Firewall traversal (through UDP hole-punching techniques). The following autonomic features make IPOP easily deployable:

a) **Self-configuration:** IPOP nodes self-configure IP tunnels to connect to other nodes on the

network, and self-configure virtual IP addresses using a DHCP implementation over a Distributed Hash Table (DHT) [6].

b) **Self-healing:** Routing in the IPOP virtual network is resilient to faults including node failures and routing outages; nodes respond to such faults in a decentralized manner by creating new edges to other nodes in the overlay.

c) **Self-optimization:** IPOP provides a mechanism to selectively establish 1-hop overlay connections between communicating nodes [3], which self-optimize the virtual network with respect to overlay link latency and bandwidth.

Closely working with deployed systems and end-users has presented several complementary research problems from different areas. These problems include (1) providing homogeneously configured and network accessible software environments on heterogeneous wide-area hosts for cross-domain collaboration and high-throughput computing (2) deploying structured P2P networks in connectivity constrained environments (3) efficient proxy discovery techniques when direct communication is not possible between IPOP nodes and (4) securing the virtual network to protect unmodified applications running over Wide-area. During my PhD research, I have been able to make significant advances on some of these problems; the proposed techniques have already been published (or are under preparation) in conferences and journals. These are highlighted in the next two sections.

### **3.1 High-throughput computing and Cross-domain collaboration**

Together with classic VM appliances for software dissemination, IPOP facilitates deployment of homogeneously configured wide-area clusters (called WOWs) [3][4] on heterogeneous hosts owned by different organization and individuals. These systems (1) scale to large number of nodes, (2) facilitate the addition of nodes through self-configuration of virtual network links, (3) maintain IP connectivity even if virtual machines migrate across subnets, and (4) present to end-users and applications an environment that is functionally identical to a local-area network or cluster of workstations. By doing so, WOW nodes can be deployed independently on different domains, and WOW distributed systems can be managed and programmed just like local-area networks, reusing unmodified subsystems such as batch schedulers, distributed file systems, and parallel application environments that are very familiar to system administrators and users. The deployment of WOWs is greatly facilitated by packaging all the software (IPOP, Condor or Globus middleware) within the VM and requiring only a NAT network, as well as by the availability of free x86-based VM monitors.

### **3.2 Structured P2P systems in connectivity constrained Wide-area environments**

At the core of IPOP architecture is Brunet P2P system that provides the structured overlay over which virtual IP traffic is tunneled and decentralized storage through a Distributed Hash Table (DHT). Deploying structured P2P systems on wide-area is a well-recognized challenge, when majority of hosts are behind NATs and Firewalls, and presence of Internet route outages (link failures, BGP routing updates, and ISP peering disputes) affect overlay structure maintenance, often leading to inconsistent routing decisions. The Brunet P2P system incorporates mechanisms for creation of overlay links between NATed nodes, through decentralized UDP hole-punching. To further self-protect the network against overlay link outages when direct communication (over TCP or UDP transports) is not possible, I have implemented a technique that facilitates structure maintenance by tunneling an overlay link through common neighbors [1]. Simulation results show this technique can improve the all-to-all routability of a 1000 node from 90% to 99%, when nodes only have a small (~ 70 %) likelihood of being able to communicate.

#### **4. Contributions**

The key contribution of my research is applying structured P2P techniques to network virtualization in Wide-area networks, thus resulting in self-configuring and decentralized system. Other network virtualization techniques (VNET [11], VIOLIN [12]) have centralized components and require an administrator to setup overlay routes and topologies, which makes these systems difficult to manage with size. IPOP is an end-to-end solution that does not require any changes to the site NATs or Firewalls configurations; it does not introduce any new network components and does not require any modifications to existing applications. This ability to operate unobtrusively and to support unmodified applications differentiates IPOP from related work addressing connectivity problems in Wide-area, which includes CODO [8] and GCB [9] from Condor Group at University of Wisconsin, and SmartSockets [10] from Virje University.

The Brunet P2P system is a unique “structured” P2P system that incorporates several mechanisms to enable structured maintenance, when connectivity is constrained due to large number of nodes behind NATs and presence of BGP route outages. With Brunet, it is possible to deploy a DHT that allows NATed nodes to join the overlay and contribute to storage, unlike systems such as OpenDHT [13] which run on a public infrastructure (e.g. PlanetLab) and nodes behind NATs can only act as DHT clients.

#### **5. Impact**

I have complemented my research with implementation of working software that not only verifies the research hypotheses, but makes it also available for peer evaluation and improvement. Together with researchers at University of Florida, I have been involved in the development of the Brunet P2P library. The Brunet P2P library is a complex software system consisting of over 40,000 lines of C# code. Months of testing on over 400 hosts on Planet Lab hosts (with highly varying load and connectivity), has helped us discover and fix bugs which included deadlocks and starvation, thus resulting in a very robust system. Both IPOP and Brunet are available under the GPL license.

The WOW techniques have resulted in two easily deployable and highly usable VM appliances (<http://www.grid-appliance.org>): (1) a Grid appliance that configures ad hoc Condor pools on Wide-area hosts for high-throughput computing, (2) and a Hadoop appliance, which enables easy creation of virtual clusters capable of running map-reduce tasks. These appliances are very useful in education for providing students with working knowledge of Condor and Map-Reduce – the use of virtual machines and networks greatly minimizes the effort required to deploy and configure these complex distributed systems on a large number of hosts. Together with researchers at IBM TJ Watson Research Center [5], I conducted a case study that demonstrated the reduction in configuration complexity of Trade6 (a Websphere based distributed application) using VM images for deployment, through the reduction in number of configurable parameters. Similarly, to secure unmodified applications on Wide-area, configuring solutions such as IPSec for different subnets and making them work through NATs is often difficult and error-prone. Virtual networks aggregate Wide-area hosts (in different subnets) into a homogeneous address space, which makes IPSec configuration very easy.

Using the Grid appliance, we have deployed a resource pool for running compute-intensive jobs through Condor. The pool consists of more than 80 compute nodes (and over 400 Planet Lab router nodes) in several NATed/Firewalled domains, and new resources can be easily added by downloading the Grid appliance image and instantiating it. This pool has been used to run over a thousand jobs from nanoHUB (<http://www.nanohub.org>), and by Ocean modelers at University of Florida for running storm-surge simulations.

The IPOP virtual network was demonstrated live at International Conference on Autonomic Computing (ICAC) 2007. A Condor worker node was instantiated on a laptop (that only had Wi-Fi connectivity provided by the conference organizers); it became part of a wide-area pool, and could submit and run Condor jobs.

## **6. Future work**

I plan to continue research on large-scale autonomous systems and to apply my experiences from areas of overlay networking and virtual machine technology to such systems. Building, maintaining, and deploying large-scale systems like IPOP and Brunet provide motivation and necessary expertise to investigate problems that target: (1) improving the IPOP virtual network – in the process develop generic solutions that are usable in other decentralized systems, and (2) explore the applicability of self-organizing virtual networks in scenarios other than Wide-area. The ensuing portions of this section describe some problems that I find worth exploring.

### **6.1 Efficient proxy discovery techniques and secure overlay**

The IPOP virtual network supports creation of direct connections between virtual IP nodes in response to communication. Situations often arise (due to certain NATs/Firewalls or BGP outages) when such direct communication is not possible. In such cases, it is possible to route communication between IPOP nodes through properly chosen proxy nodes. This problem of proxy discovery is a very generic problem that also applies to other wide-area applications. The choice of proxy depends on the goals of application; a VoIP application may be more interested in minimizing latency, whereas a bulk data transfer application is more concerned with maximizing bandwidth. Current systems such as Skype use a centralized directory to track information of potential proxy nodes, which are referred to as “supernodes”. However, in a system such as IPOP/WOW where resource owners are different organizations or individuals, maintaining such central servers is not easy. To discover suitable proxies in a completely decentralized manner, my approach would be to apply synthetic coordinates (such as Vivaldi), in conjunction with discovery based on P2P search techniques.

The Grid appliance currently uses IPSec to secure communication between virtual IP nodes. To further protect the underlying P2P overlay from malicious hosts, I want to investigate the feasibility of using Datagram TLS (UDP-based) to only allow authorized hosts to join the overlay; at the same time not compromising the current ability to work across NATs (SSL/TLS work with TCP that is not amenable to NAT-traversal).

### **6.2 Network virtualization in data center environments**

Network virtualization techniques are also applicable in data center running third-party applications. Virtual machines (VMs) are already popular in these scenarios because of their ability to sandbox execution of an untrusted application, as in Amazon’s Elastic Compute Cloud. Complementary to VMs, virtual networks isolate the traffic generated by such applications, thus preventing any damage to non-participating hosts from a malicious user application that would otherwise share the same network. In addition, there are scenarios where not enough IP addresses are available for assignment to VMs; in such cases virtual networks allow VMs running on different hosts to be able to communicate without requiring a physical IP address on the LAN.

Facilitating the use of virtual networks in these scenarios would require (1) developing services for dynamic provisioning of such isolated virtual networks for different users, (2) mechanisms to express and enforce QoS guarantees, and (3) if necessary, kernel and VMM enhancements to minimize the overhead of network virtualization for high-end applications.

### **6.3 Virtual networks of VM appliances**

Virtual machines allow migration of unmodified applications because of the decoupling from the underlying host environment they provide. Furthermore, with the decoupling of IP address space provided by virtual networks, it is even possible to migrate an entire distributed system (a multi-tier application or a part of it) to a new site without any reconfiguration, thus facilitating quick recovery from failures by greatly reducing the redeployment effort.

### **6.4 Decentralized network services**

The IPOP virtual network uses a decentralized implementation of the DHCP protocol based on DHT, for virtual IP configuration of hosts. I want to investigate the applicability of P2P techniques to improve scalability, reliability and manageability of other network services and information systems.

### **References**

[1] A. Ganguly, D. Wolinsky, P. Boykin, R. Figueiredo. Improving correctness of structured routing in connectivity constrained wide-area environments". in preparation.

[2] A. Ganguly, A. Agrawal, P. Boykin and R. Figueiredo. IP over P2P: Enabling Self-Configuring Virtual IP Networks for Grid Computing. In Proceedings of Parallel & Distributed Processing Symposium (IPDPS), Apr 2006.

[3] A. Ganguly, A. Agrawal, P. Boykin, R. Figueiredo. WOW: Self-Organizing Wide Area Overlay Networks of Virtual Workstations. In Proceedings of High Performance Distributed Computing Symposium (HPDC), Jun 2006.

[4] D. Wolinsky, A. Agrawal, P. Boykin, J. Davis, A. Ganguly, V. Paramygin, P. Sheng, R. Figueiredo. On the design of Virtual Machine Sandboxes for Distributed Computing in Wide Area Overlays of Virtual Workstations. In Workshop on Virtualization Technologies in Distributed Computing (VTDC), Nov 2006.

[5] A. Ganguly, J. Yin, H. Shaikh, D. Chess, T. Eilam, R. Figueiredo, J. Hanson, A. Mohindra, G. Pacifici. Reducing Complexity of Software Deployment with Delta Configuration. Short-Paper at Symposium on Integrated Network Management (IM), May 2007.

[6] A. Ganguly, D. Wolinsky, P. Boykin and R. Figueiredo. Decentralized dynamic host configuration in Wide Area Overlays of Virtual Workstations. In Workshop on Large-scale and Volatile Desktop Grids (PCGrid), March 2007.

[7] Ivan V. Krsul, A. Ganguly, J. Zhang, Jose A. B. Fortes, Renato J. Figueiredo. VMPlants: Providing and Managing Virtual Machine Execution Environments for Grid Computing. In Proceedings of Supercomputing Conference, Nov 2004.

### **Related References**

[8] S. Son, B. Allcock, M. Livny. CODO: firewall traversal by cooperative on-demand opening. In Proceedings of International High Performance Distributed Computing Symposium (HPDC), Jul 2005.

[9] S. Son, M. Livny. Recovering Internet Symmetry in Distributed Computing. In Proceedings of Cluster Computing and the Grid Symposium, May 2003.

[10] Jason Maassen, Henri E. Bal. Smartsockets: Solving the connectivity problems in grid computing. In Proceedings of High Performance Distributed Computing Symposium (HPDC), Jun 2007.

[11] A. Sundararaj, P. Dinda, Towards Virtual Networks for Virtual Machine Grid Computing. In Proceedings of Virtual Machine Research and Technology Symposium (VM 04), May 2004.

[12] X. Jiang and D. Xu. Violin: Virtual internetworking on overlay infrastructure. In Proceedings of Symposium on Parallel and Distributed Processing and Applications, Dec 2004.

[13] Sean Rhea, Brighten Godfrey, Brad Karp, John Kubiatowicz, Sylvia Ratnasamy, Scott Shenker, Ion Stoica, and Harlan Yu. OpenDHT: A Public DHT Service and Its Uses. Proceedings of ACM SIGCOMM, August 2005.