Statement of Research Interest

Shah Asaduzzaman

Research Area and Approach

My primary research interest is in the area of networked and distributed systems. Specifically, I work on the systems of interconnected resources where heterogeneity, unreliability and uncertainty of state information are omnipresent. My general research question is how system architecture and resource management algorithms can support emergent system characteristics such as quality assurance, availability and resource efficiency.

Dissertation and Current Research

With the presence of planetary scale communication platform like Internet, different types of resources can now be shared among a widespread community for various types of computing applications. Architectures like *Grid computing* systems promise acquisition of computing power for large scientific computations from a collaborative resource sharing platform. *Utility computing* systems are envisioned where server farms are established to allow businesses to acquire computing resources on-demand instead of self provisioning. So far implementations of such utility services have mostly relied on clusters of servers or Internet data centers (IDCs). These systems are statically dimensioned which means accurate resource requirement analysis should be done on the services before provisioning them. Accurate requirement estimation is especially a hard problem for services with geographically dispersed users. Static installations with over-provisioning may overcome the problem in some cases, but for services that have variable resource demands, dynamic provisioning is a necessity.

In my PhD dissertation, I introduced a new service deployment architecture named Public Computing Utility [5, 6] that combines statically provisioned dedicated resource pools with widely available opportunistic public resources to provide quality assured services. The central idea of my proposed architecture is to classify the globally available resources into dedicated and opportunistic public resources. Several scientific public computing projects in various disciplines such as physics, meteorology and bioinformatics has already proved the huge potential of idle capacities of desktop computers. I have shown that quality assured value added computing services can be created by dynamically allocating resources from a combination of dedicated and public resource pools and maximize the overall system throughput and revenue of the utility service provider. My main objective was to investigate the design parameters and trade-offs of several possible architectural models of the resource management systems through extensive simulation studies and to devise appropriate scheduling algorithms for each of the models.

One major overhead of large-scale distributed systems is the aggregation of global status information, especially with highly dynamic components like opportunistic resources. To avoid such overhead, I first developed a state-oblivious resource management model for the computing utility and devised scheduling algorithms to achieve high throughput and compliance to service level agreements [2, 5].

In another model, I designed the resource manager to receive periodic status updates from remote public resources and revise the scheduling decisions with new information [1, 4]. I also considered the possibility of live migrations on jobs between computers over the network according to revised decisions. Both status updates and migrations cause significant communication overhead. Through simulation studies I addressed the research questions of how much benefit such informed and dynamic scheduling can achieve in terms of customer compliance and revenue generation, and how the communication overhead restricts the scalability of the system.

Finally, I applied the bi-modal approach on resource management for distributed stream processing applications. Stream processing model is useful for several interesting applications such as processing continual queries on sensor data-streams and transcoding of multimedia data streams. In addition to computational resources, proper allocation of network bandwidth is crucial for such applications. The proposed stream processing system connects a set of application servers with a combination of dedicated leased lines and public ip network, and I developed resource management algorithms that can dynamically allocate server and network resources to maximize number of successful streaming sessions across the system [3].

Future Research Directions

I envision my future research to span across some interrelated sub-areas of computer science, such as service oriented architectures, distributed computing systems and ubiquitous computing. The following paragraphs give some abstract descriptions of the contexts of my future research projects.

Energy Aware Hosting Platforms

On one hand energy is becoming the most precious resource of upcoming decades; on the other hand uncontrolled usage of energy is causing environmental disasters threatening the very survival of human species. My vision is that resource collaboration among hosting platforms and incorporation of public resources can be leveraged to develop an energy efficient architecture for service hosting. Resource scheduling algorithms need to be developed that considers energy as a first class resource in the system. Several issues need to be addressed to realize such platforms including, a) secure and lightweight isolation mechanism for guest applications b) resource management algorithms that maximize utilization of all types of resources including energy, avoid unnecessary recruitment of resources and able to co-allocates different resources such as processing bandwidth, storage capacity, and network bandwidth. I believe solving these problems will enable new generations of hosting platforms and also new classes of applications.

Service Composition in Ubiquitous Computing Environment

Rapid advancement in hardware technology for mobile and small scale smart devices has prompted scientists to realize the visions for a ubiquitous or pervasive environment of computing. An important dimension of the ubiquitous computing environments is their extensibility. On demand creation of new personalized services from available components is a necessity in these environments due to the overwhelming variety in service demands. Although the service composition problem has been studied widely in the realm of web based services, the ad-hoc nature of ubiquitous computing demands dramatically different approaches. Important research issues include discovery and recruitment of services in absence of any global state information, decentralized and cooperative middleware for the composition process, dealing with the heterogeneity of devices and network capacities, finding services with appropriate trust relationships with the requester, adaptive maintenance of the composed service in presence of component failures, etc.

Distributed Stream Processing Architectures

An emerging set of applications require on-line multi-level processing of a stream of data generated by some source in the network and deliver the resultant stream to the user [7]. Some example applications include sensor networks, media multiplexing and trans-coding, location-tracking services, fabrication line management, and network management. These applications are characterized by the need to process high-volume data streams in a timely and responsive fashion. Such stream processing also enables parallel or pipelined processing large volume of data that some scientific research projects need to process. A federated resource management architecture is needed for scalable, load-resilient usage of in-network resources for stream processing. Several issues such as allocation of computing resources constrained by network quality properties, appropriate ways to deal with resource heterogeneity, network congestions, load balancing, etc. need to be addressed. Opportunistic usage of public resources may serve towards a scheme that is tolerant of highly varying loads in the streams. Also, decentralized resource management algorithms need to be developed to avoid status monitoring of huge number of small resource nodes.

Summary

In summary, my research plan for near future will be motivated by my findings during the PhD research as well as my visions for advancement of the human society.

Reference

1. S. Asaduzzaman and M. Maheswaran, “Strategies to Create Platforms for Differentiated Services from Dedicated and Opportunistic Resources”, *J. Parallel and Distributed Computing*, to appear.
2. S. Asaduzzaman and M. Maheswaran, “Utilizing Unreliable Public Resources for Higher Profit and Better SLA Compliance in Computing Utilities” , *J. Parallel and Distributed Computing*, 66(6), pp. 796-806, 2006
3. S. Asaduzzaman and M. Maheswaran, “Towards a Decentralized Algorithm for Mapping Network and Computational Resources for Distributed Data-Flow Computations”, *21st IEEE Intl. Symp. on High Performance Computing Systems and Applications–HPCS-2007*, May 2007, Saskatoon, SK, Canada. to appear.
4. S. Asaduzzaman and M. Maheswaran, “Heuristics for Scheduling Virtual Machines for Improving QoS in Public Computing Utilities”, *9th Intl. Conf. on Computer and Information Technology –ICCIT-2006*, Dec. 2006, Dhaka, Bangladesh.
5. S. Asaduzzaman and M. Maheswaran, “Leveraging Public Resource Pools to Improve the Service Compliances of Computing Utilities”, *Springer LNCS(3296): proceedings of Intl. conference on High-Performance Computing (HiPC)*, pp. 242-251, Dec. 2004.
6. M. Maheswaran, B. Maniymaran, S. Asaduzzaman, and A. Mitra, “Towards a Quality of Service Aware Public Computing Utility”, *Adaptive Grid Computing Workshop*, Cambridge, MA, Aug. 2004
7. P. Pietzuch et al. “Network-Aware Operator Placement for Stream-Processing Systems”, *22nd International Conference on Data Engineering (ICDE'06)*, Apr. 2006.