# Position Paper:
# Verification of Data Intensive High Performance Computing Middleware

**Venkatram Vishwanath**
venkat@evl.uic.edu
University of Illinois Chicago

**Lenore Zuck**
lenore@cs.uic.edu
University of Illinois Chicago

## Data Intensive High Performance Computing:

Interactive exploration of terabyte and petabyte datasets has been identified as a critical enabler for scientists to glean new insights in a variety of disciplines, such as biomedical imaging, geosciences and high-energy physics [1]. Practically, these large-scale datasets must flow among a Grid of instruments, physical storage devices, visualization displays, and computational clusters. These applications have a real need for tens to hundreds of gigabits-per-second of bandwidth and deterministic QoS that are best satisfied by interconnecting Grid resources with dedicated networks dynamically created by concatenating optical lightpaths (lambdas). This is called a LambdaGrid [2]. LambdaGrids are routinely used in multiple domains, including, real-time weather prediction and forecasting, interactive remote visualization, tropical cyclone analysis, Metagenomics, distributed data mining and astronomy. LambdaGrid applications typically use middleware, including, GLOBUS, CACTUS. LambdaRAM [3], a multi-dimensional distributed data middleware encompassing multiple clusters interconnected via ultra high-speed optical networks, has been demonstrated with geoscience and bioscience applications. NASA is currently working towards the integration of LambdaRAM for distributed data analysis for real-time weather prediction [4]. Data Intensive middleware typically use MPI, traditional and high performance sockets for intra-cluster communication and specialized transport protocols for wide-area communication, including High-speed variants of TCP, RBUDP, and UDT.

## Architectural Trends in High Performance Computing (HPC) systems

Multicore architectures have established themselves as an integral part in HPC systems. Quad-core architectures, from Intel and AMD, and, dual-core processors from IBM, are gaining prominence in HPC clusters. With the introduction of Terascale chips from Intel, IBM and AMD, with a roadmap for usage in desktop and server environments, one would expect these to be routinely used for advanced e-science applications in the near future. Additionally, these

processors incorporate heterogeneous cores, including, graphical processing units (GPU) and hardware accelerators. NUMA, SMP and their hybrid combinations, together with a multi-dimensional topology, characterize the memory architecture of these future processors.
High performance networks such as 10G Ethernet, Myrinet, Infiniband, are an integral part of these system for performance and scalability. 40G and 100G Ethernet have been recently standardized with a strong recommendation to employ Dense Wavelength Division Multiplexing (DWDM) technology in their design. 40G network interface cards, designed as 4 x 10G DWDM based NIC, are being tested in research labs with the goal for deployment in HPC systems. We are also witnessing the advent of on-chip optical interconnects that overcome the bandwidth and power limitations of current electrical system interconnects. These above will form the fundamental building block for upcoming Petascale and Exascale architectures. However, novel techniques are required to enable applications and data-intensive middleware to reliably utilize the potential of these cyber-infrastructures.

**Advocated Approach (Position):**

Currently, much (if not all) HPC middleware is designed in an ad-hoc manner. Verification of these middleware is usually accomplished after an initial implementation. Formal verification of a complex and large code base, is, however, virtually impossible. In our work towards the initial specification and verification of LambdaRAM [5], we decomposed the existing implementation into components, abstracted the components, and only then succeeded in formally verifying the code against the abstractions. Additionally, These abstractions have also been useful in the extension of LambdaRAM. We strongly believe that decomposition of existing middleware, to identify reusable building blocks, would benefit the modeling and verification of such complex systems. We believe that using LambdaRAM as a use-case would be an excellent initial step towards the reusable building blocks. In the course of the verification of LambdaRAM, we found that incorporating locality in order to exploit compositionality, including intra-node communication, intra-cluster communication, communication among peer-clusters, was extremely helpful. Similar locality-aware compositional techniques would also benefit HPC middleware such as Partitioned Global Address Spaces (PGAS) and Global Arrays (GA). Key features of the data intensive middleware and data-intensive cyber-infrastructures, include, variable network latencies, message re-orderings and highly parameterized features. We believe that our previous work in verification of parameterized systems [6-9] can aid in the verification of the data-intensive HPC middleware. We firmly believe that one needs to build a tool-suite incorporating model checking and theorem proving for the verification Data-intensive HPC systems. We also believe that to improve the adoption of verification in HPC, tools with intuitive user interfaces and scripting capabilities are essential.

## References:

[1] J. Leigh, L. Renambot et al, "The Global Lambda Visualization Facility: An International Ultra-High-Definition Wide-Area Visualization Collaboratory", Journal of Future Generation Computer Systems (FGCS) Vol 22 (2006).

[2] X. Wang, V. Vishwanath, B. Jeong, R. Jagodic, E. He, L. Renambot, A. Johnson, J. Leigh, LambdaBridge: A Scalable Architecture for Future Generation Terabit Applications. Broadnets 2006 - San Jose, CA, 10/01/2006 - 10/05/2006.

[3] Krishnaprasad, N., Vishwanath, V., Venkataraman, S., Rao, A., Renambot, L., Leigh, J., Johnson, A., JuxtaView – A Toold for Interactive Visualization of Large Imagery of Scalable Tiled Displays. Proceedings of IEEE Cluster 2004, San Diego, CA, 09/20/2004 - 09/23/2004

[4] V. Vishwanath, R. Burns, J. Leigh, M. Seablom, Accelerating Tropical Cyclone Analysis using LambdaRAM, To appear in Journal of Grid Computing

[5] V. Vishwanath, L. Zuck, J Leigh. Specification and Verification of LambdaRAM – A Wide Area Distributed Cache for High Performance Computing. To appear in ACM IEEE MEMOCODE 2008

[6] A. Pnueli, S. Ruah, and L. Zuck. Automatic deductive verification with invisible invariants. In Proc. 7th Intl. Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'01), volume 2031 of Lect. Notes in Comp. Sci., Springer-Verlag, pages 82--97, 2001.

[7] Y. Fang, N. Piterman, A. Pnueli, and L. Zuck. Liveness with incomprehensible ranking. In Proc. 10th Intl. Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'04), volume 2988 of Lect. Notes in Comp. Sci., Springer-Verlag, pages 482--496, April 2004.

[8] Y. Fang, N. Piterman, A. Pnueli, and L. Zuck. Liveness with invisible ranking. In Proc.of the 5th conference on Verification, Model Checking, and Abstract Interpretation, volume 2937 of Lect. Notes in Comp. Sci., pages 223--238, Venice, Italy, January 2004. Springer-Verlag.

[9] L. Zuck and A. Pnueli. Model checking and abstraction to the aid of parameterized systems. Computer Languages, Systems, and Structures, Volume 30(3--4) pp.139--169 2004.