

MeDICI: Middleware for Data-Intensive Computing

Challenge:

More than ever, technological advancements are producing massive amounts of data from high-throughput instrumentation, sophisticated system sensors, and modeling and simulation programs. This deluge of complex, high-volume data burdens scientists and analysts working to make sense of the information and its relationship to intricate problems. Building analytical software systems that can process this data in a timely fashion presents challenges in many ways, including:

- Capturing and integrating high-throughput data from its source.
- Integrating multiple algorithms for fusing and analyzing data in real time.
- Managing diverse data formats and distributed data sources.
- Integrating distributed, heterogeneous software and hardware systems into a single application.

In our data-intensive research program, scientists at Pacific Northwest National Laboratory (PNNL) are working to create new technologies to solve these challenges. At the core of these emerging technologies is the Middleware for Data-Intensive Computing (MeDICI) Integration Framework, an integration middleware platform designed to solve data analysis and processing needs of scientists across many domains, and in a fashion that is scalable, easily modified, and robust to multiple languages, protocols, and hardware platforms.

Capability:

MeDICI is designed for building complex, high-performance analytical applications, typically comprising a pipeline of software components. Each component in the pipeline performs some analysis on incoming data and transfers the results to the next step(s) in the pipeline. The MeDICI framework enables software codes written in any languages to be wrapped as MeDICI components, which can then be simply plugged together using the core framework to create applications. The framework automatically takes care of difficult tasks such as multi-threading, message buffering, distributed communications, and load balancing. This makes it simple to integrate separate codes, which were not designed to work together, into complex applications that operate as a data analysis pipeline.

MeDICI was designed to address many of the difficult aspects of building analytical applications, namely:

Pipeline creation – MeDICI makes it easy to transfer data as it moves from one application to another, turning a set of distributed heterogeneous components into an integrated pipeline.

Handling large data – MeDICI offers features that give pipeline designers choices on how to pass data through pipelines to maximize the performance of the applications.

Component libraries – MeDICI enables analytical codes written in any language and running on any platform to be plugged into a MeDICI pipeline through the creation of a few lines of code and without changing the analysis code itself. The codes, in fact, are oblivious to the MeDICI framework, making it easy to combine existing components in an application.

Impact:

MeDICI provides a faster, more efficient, and less expensive way to build analytical pipelines. The platform provides a scalable, flexible, and extensible development and run-time framework for ease of use in many application domains.

Applications:

MeDICI is being applied to a variety of research projects at PNNL, including:

Bioinformatics Resource Manager: Data-intensive pipelines that analyze large biological data sets are executed and managed on a 32-node cluster using MeDICI.

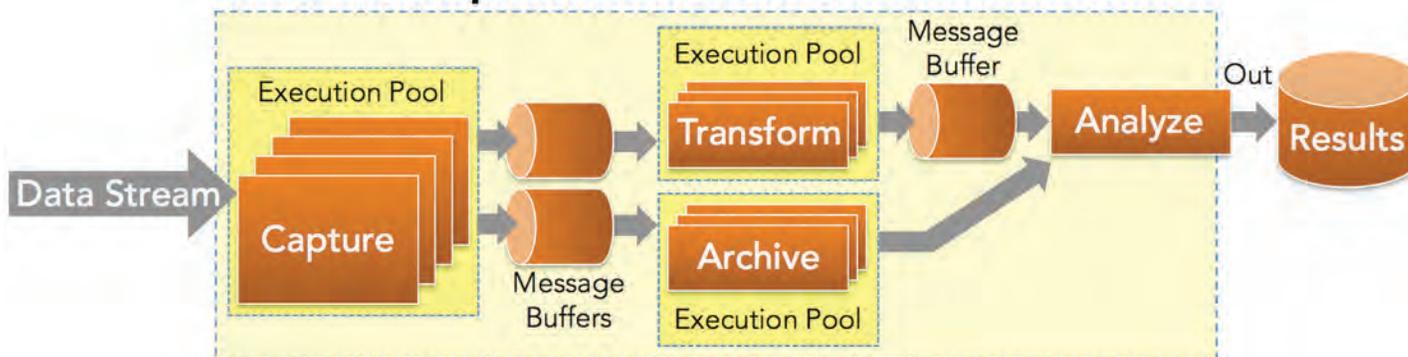
Power Grid Failure Analysis: A high-performance MeDICI pipeline sends data from a Cray multithreaded supercomputer to a conventional supercomputer cluster for simulation. These results, on the order of 100's of MBs, are returned by the pipeline to the Cray for detailed analysis.

Text Analysis: MeDICI integrates a set of components that perform advanced semantic text analysis into a high-throughput processing pipeline.

MeDICi Pipeline - Design



MeDICi Pipeline - Runtime



MeDICi Features:

Simple. The MeDICi framework automatically handles many of the complex architectural issues that must be addressed when building high-performance software pipelines.

Robust. MeDICi is built on proven standards-based integration, workflow, and provenance technologies.

Flexible. MeDICi supports multiple languages, communication protocols, and hardware platforms.

Efficient. MeDICi improves performance by passing large data by reference.

Publications:

Gorton I, CS Oehmen, and JE McDermott. 2008. "It Takes Glue to Tango: MeDICi integration framework creates data-intensive computing pipeline." *Scientific Computing*. 25(7):16-24.

Gorton I, AS Wynne, JP Almquist, and J Chatterton. 2008. "The MeDICi Integration Framework: A Platform for High Performance Data Streaming Applications." In *WICSA 2008. 7th IEEE/IFIP Working Conference on Software Architecture*, pp. 95-104. February 18-22, 2008, Vancouver, Canada. IEEE Computer Society, Los Alamitos, CA.

Gorton I, JM Chase, AS Wynne, and JP Almquist. 2009. "Services + Components = Data Intensive Scientific Workflow Applications with MeDICi." In *12th International Symposium on Component Based Software Engineering (CBSE 2009)*, pp. 227-241, June 2009, Springer-Verlag.

Chase JM, I Gorton, C Sivaramakrishnan, JP Almquist, AS Wynne, G Chin, and TJ Critchlow. 2009. "Kepler + MeDICi - Service-Oriented Scientific Workflow Applications." In *IEEE 2009 International Conference on Scientific Workflows*, pp. 275-282. July 6-10, 2009, IEEE Congress on Services, Los Angeles, California. IEEE Computer Society, Los Alamitos, California.

SC 08 Award:

MeDICi was a core feature in the 2008 award captured by PNNL for "Best Overall" Supercomputing (SCO8) HPC Analytics Challenge. PNNL's entry, directed at genomics, combined multiple databases, analysis software, and the PNNL-developed visualization technology called *Starlight*, which presents data in unique visual patterns and allows users to interactively explore them. PNNL addressed a common problem for genomics researchers, where desktop computers frequently are not able to handle the volume of calculations needed to analyze many genomes at once. Using the MeDICi middleware, PNNL was able to demonstrate an iterative workflow that effectively pulls together data, analysis, and visualization.

- To learn more about MeDICi and instructions for a free download go to: <http://medici.pnl.gov>

Contact:

Dr. Ian Gorton, the designer of MeDICi, is a Senior Research Scientist for the Computational Sciences and Mathematics Division at PNNL and is the Chief Architect for PNNL's Data Intensive Computing Initiative. He leads research projects and consults across the Laboratory on the creation of new methods and technologies for building complex software systems. Dr. Gorton has written two books on software architecture, co-authored 30 referred journal articles and 80 referred international conference papers, and has presented tutorials at major software engineering conferences. He currently is working on the second edition of his book, *Essential Software Architecture*, which will include coverage of MeDICi.



Ian Gorton, Chief Architect

Ian.gorton@pnl.gov
(509) 375-3850