Annual Progress Report – Year 2, Statement of the unexpended funds at the end of the budget period, FA-TC-0018.2 Continuation Application

Project Title: **TRALEIKA GLACIER X-STACK**
PI: **Shekhar Borkar**
Cooperative Agreement #: **DE-FC02-12ER26095**
Award #: **DE-SC0008717**
Recipient: **Intel Federal LLC**
Period Covered by Report: **September 1, 2013 to May 31, 2014**
Report Date: **May 30, 2014**

Acknowledgment: This material is based upon work supported by the Department of Energy [Office of Science] under Award Number DE-SC0008717.

Table of Contents

Progress/accomplishments during the current funding period and plans for the next year funding period

Executive Summary

The straw-man Exascale architecture is stable, captured in a functional simulator (FSim) used to evaluate the software stack. The programming system components, concurrent collections, hierarchical tiled arrays, and the R-Stream optimizer work together, demonstrated by porting LULESH (a refactored DOE applications) on the simulator, and more in progress. The open community runtime is released to the community for researching low level system SW and execution models. Preliminary demonstration of the entire SW stack, running a few key refactored application, by August 2014 is on track.

Progress, Accomplishments, and Status

We are shifting our efforts from infrastructure building—major thrust last year—to researching the X-Stack software using the infrastructure.

We have identified the Conjugate Gradient (CG) computation as a key kernel, being relevant for both CESAR (CG is dominant in the Nekbone proxy app) and ExMaT co-design center's apps, and have started the investigations. We are also focusing on the CoMD proxy application, and on track to demonstrate refactored Lulesh on the research software stack in the upcoming applications workshop (formerly Hackathon).

Porting of OCR (the open community runtime) to our simulator (FSim) has started with the aggressive goal of getting it ready for the workshop by mid-January. OCR was also highlighted at the Birds of a Feather session at SC13, with demonstration of the OCR runtime running the Unbalanced Tree Search benchmark. The novelty is in adaptation to changes in the environment, such as cores going offline or runtime goals changing (high performance to low power consumption, etc).

The programming system is making good progress. The CnC (Concurrent Collections) tuning system is working, provides 19% improvement on an unturned asynchronous execution on a shallow platform with today's costs. We expect more benefits on a deeper system or with the future costs. PNNL worked on a high-level representation of Lulesh in CnC. Furthermore, we have also implemented some of the NAS benchmarks in HTAs (Hierarchical Tiled Arrays).

We introduced ISA modifications to more compactly encode integer signed and unsigned arithmetic, saving about 50 instructions, or about 10% of the opcode space. These changes were reflected by updating the LLVM compiler and binutils to keep the tool chain current. We also continue to integrate the APIs for power and for software managed caches with the compiler and the FSim simulator.

To investigate introspection and self-awareness, we have added capabilities and infrastructure to the FSim simulator required for basic feedback loops needed for management algorithms. These include counters and registers in the hardware description, as well as capability of automatically simulating heat generation and heat transfer for any chip configuration and ambient temperature. We expect demonstration of a simple feedback loop in the near future.

To improve tuning of the system, we now have a visualization tool that allows viewing of energy usage throughout the system. This tool can display energy usage in a variety of ways, allowing filtering by EDTs, by hardware component, or by time. As an example, tracking data movement in OCR has allowed us to generate datasets that can be fed into this visualization tool to demonstrate the effects of local memory size on total energy usage.

An application Workshop brought together the members of the Traleika Glacier project as well as the co-design centers for three days. We discussed current progress and got much better understanding of the science behind the applications of interest to DoE. Specifically, we discussed the work on LULESH and started exploring combustion and its adaptive mesh refinement (AMR) and multi-grid (mini GMG) components. As a result, the applications team is now focused on two applications to port on OCR and FSim: Conjugate Gradient (CG-kernel) and CoMD, both are working on OCR-x86 and will be targeted to FSim.

The open community runtime (OCR) port on an x86 platform is very stable. The OCR port on TG architecture now runs on an x86 platform for quick validation and debug; we are making steady progress to run it on the TG architecture simulator (FSim) this quarter.

The specification for a new 64 bit ISA is released to the wider Traleika-Glacier team, which incorporates feedback from applications, compiler development, runtime, and hardware. Specific key features include: 64 bit encoding, support for transcendental functions, enhanced DMA and hardware queue operations, and interrupt capabilities. To evaluate the architecture with TG's software stack, ETI focused on improving the simulator to match the changes. Their data movement model shows that although memory size does have a large effect on data movement energy, it is overshadowed by the constant leakage energy, with little effect on the total energy dissipation of the system.

To support 64 bit transition, we have updated binutils, and ongoing compiler support. We have started our efforts to apply R-Stream for generating optimized OCR versions of the proxy application programs, and understanding the miniGMG application benchmark to facilitate the mapping of miniGMG through R-Stream to generate an optimized OCR version of the miniGMG code.

Our high-level programming model, CnC, and our low-level programming model, OCR, are very consistent in their asynchronous, event-driven, task-based execution. Transition of the user whiteboard version of LULESH to a CnC graph specification was straightforward. We both validated the ease of use of CnC and also uncovered some new optimization potential. The Hierarchical Tiled Array (HTA) work shows potential for programming productivity with performance matching that of the tuned OpenMP code.

University of Delaware team demonstrated the ability to model heat dissipation and transfer in a multi-block simulation (9-blocks) with the capability of the simulator (FSim) to trace energy and heat. The demonstration was performed using a matrix-vector kernel as a basis. This is a start to incorporate introspection and self-awareness in the system software.

On the architecture front, we have continued to integrate the APIs for energy management and software managed caches in the Fsim simulator and automate their use with the compiler to demonstrate that the compiler can generate code for these API automatically and efficiently.

Plans for the next year

Having built the infrastructure first year, followed by setting the technology direction the second year, our thrust next year will be to stabilize and mature the technologies so that the community can take it further. The straw-man architecture is stable and will be frozen, captured in the latest version of the simulator incorporating rudimentary interconnect architecture. The programming system components consisting of CnC, HTA, and R-Stream will be integrated into a coherent system. The open community runtime will run on the TG architecture (FSim) to help make architecture tradeoffs. And the low level system software will incorporate introspection and self-awareness. The entire TG software stack will be used to run key refactored DOE applications to establish both the Exascale system architecture and the software stack to effectively exploit the architecture—in the true spirit of HW/SW co-design, partnering with our co-design partners.

Schedule Status

Our progress, accomplishments, and status above represents milestones 5 through 7 below. We consider ourselves on track overall. We employ Shekhar Borkar's PI leadership, hold regular weekly PI and technical meetings, monthly rolling wave milestone planning meetings, semi-annual application workshops, a collaboration wiki, and central code repository to keep the team focused on priorities to achieve a successful X-Stack.

#	Due	Milestone
1	11/30/12	Architecture V2 spec & preliminary apps kernel identified for evaluation
2	3/1/13	Simulators V2 functional, tools (C + binutils) in place, IRR V1 identified
3	5/31/13	Selected kernels evaluated for 0(compute)
4	8/30/13	Basic timing in simulator, intelligent scheduling in Exec model, tools (LLVM, etc)
5	11/27/13	Selected kernels evaluated for 0(com), select apps coded with PGM system for IRR
6	2/28/14	Architecture V2.5 spec, system evaluation of V2.0
7	5/30/14	Simulators V2.5 functional, tools for V2.5 released
8	8/29/14	System evaluation of V2.5
9	11/26/14	Arch V3.0 spec, selected apps evaluation with Exec model & PGM system for V2.5
10	2/27/15	Simulators V3.0 functional, tools for V3.0 released
11	5/29/15	Release OCR (Open Collaboration Runtime) V1.0
12	8/28/15	Evaluation of all X-Stack technologies and report

Changes in approach or aims and reasons for change

Our technical approach is solid, well vetted in the community during recent X-Stack meetings, and encouraged by the community. So far the results of our work have been encouraging as well, and hence we do not envision any major changes in our technical approach, statement of work, or milestones.

As always, we encounter many challenges on the way, and have to do course correction with minor detours, especially to change priorities of the tasks. Examples of course corrections taken or underway include:

- We noticed that distributed OCR is more important than CnC tuning, and hence adjusted priorities accordingly, without changing the statement of work, milestones, or deliverables.
- We've asked Rice University to take leadership responsibility for OCR on FSim, which may result in additional funding to Rice, managed within overall existing budget.
- With co-design center responsibility removed from the award, we engaged the services of Roger Golliver under UIUC to drive hero programming of LULESH and AMR/miniGMG, managed within overall budget.
- We introduced the Modelado Foundation to the X-Stack community to enable centralization and productization of refactored open source application codelets, compilers, tools, and runtimes for X-Stack, and we gave them responsibility for the X-Stack wiki – http://xstackwiki.modelado.org, managed within overall existing budget.
- We removed Reservoir Labs' SOW item related to graph based numerical solvers. The change to the SOW was to make room for continued LLVM support that Reservoir Labs has been doing.

Actual or anticipated problems or delays and actions taken or planned to resolve them

At present, we do not envision any major problems causing delays, other than what is covered in the next section below. There will be technical challenges that will cause changes in tactics, as expected in any research, but none that will impact overall technical direction, milestones, or deliverables.

Any absence or changes of key personnel or changes in consortium/teaming arrangement

[redacted]

Descriptions of any product produced or technology transfer activities accomplished during this reporting period

Publications

The following were presented at the CnC'13 workshop September, 2013. This was the fifth annual CnC workshop. It was co-located with Languages and Compilers for Parallel Systems (LCPC) in Santa Clara, CA.

- "Compiler Optimization of an Application-Specific Runtime". Kathleen Knobe (Intel) and Zoran Budimlic (Rice)*.
- "The CnC tuning capability", Sanjay Chatterjee (Rice), Zoran Budimlic (Rice), Vivek Sarkar (Rice), Kathleen Knobe (Intel).
- "Automatic Selection of Distribution Functions for Distributed CnC", Kamal Sharma (Rice), Kathleen Knobe (Intel), Frank Schlimbach (Intel), Vivek Sarkar (Rice)*.
- "CnC on Open Community Runtime", Alina Sbirlea (Rice) and Zoran Budimlic (Rice).
- "Bounded Memory Scheduling of CnC Programs", Dragos Sbirlea (Rice), Zoran Budimlic (Rice) and Vivek Sarkar (Rice). *
- "CDSC-GL: A CnC-inspired Graph Language", Zoran Budimlic (Rice), Jason Cong (UCLA), Zhou Li (UCLA), Louis-Noel Pouchet (UCLA), Vivek Sarkar (Rice), Alina Sbirlea (Rice), Mo Xu (UCLA), Pen Zhang (UCLA).*
- "Implementing Asynchronous Checkpoint/Restart for CnC", Nick Vrvilo and Vivek Sarkar (Rice University) Kath Knobe and Frank Schlimbach(Intel)
- "Automatic CnC generation from a sequential specification", Nicolas Vasilache (Reservoir Labs, Inc.)

  Note: Asterisked (*) presentations are supportive of the Traleika Glacier X-Stack strategic aims and objectives but not directly under the statement of work.

Reservoir Labs submitted a paper for publication to PPoPP; unfortunately, this contribution was rejected. We will produce a technical report taking into account the reviewers' comments. We circulated this publication internally to members of the OCR core team and to members from Intel.

"Compiler Support for Software Cache Coherence"
Submitted for publication
Sanket Tavarageri, Wooil Kim, Josep Torrellas, and P Sadayappan Pacific Northwest National Labs (John Feo, Andres Marquez)

"T2: ASAFESSS: A Scheduler-Driven Adaptive Framework for Extreme Scale Software Stacks"
4th International Workshop on Adaptive Self-tuning Computing Systems 2014, Vienna Austria. (Best paper award)
St. John, T. et al.

"ACDT: Architected Composite Data Types Trading-in Unfettered Data Access for Improved Execution"
submitted to the 23rd International ACM symposium on High Performance Parallel and Distributed Computing 2014, Vancouver Canada
Marquez, A. et.al

"A Dynamic Schema to increase performance in Many-core Architectures through Percolation operations"
In Proceedings of the 2013 IEEE International Conference on High Performance Computing (HiPC 2013), Hyderabad, India, December 18 - 21, 2013.
Elkin Garcia, Daniel Orozco, Rishi Khan, Ioannis Venetis, Kelly Livingston, and Guang Gao.

"Optimizing the LU Factorization for Energy Efficiency on a Many-Core Architecture"
In Proceedings of the 26th International Workshop on Languages and Compilers for Parallel Computing (LCPC 2013), Santa Clara, CA, September 25-27, 2013.
Elkin Garcia, Jaime Arteaga, Robert Pavel, and Guang R. Gao.

"ASAFESSS: A Scheduler-driven Adaptive Framework for Extreme Scale Software Stacks"
In Proceedings of the 4th International Workshop on Adaptive Self-Tuning Computing Systems (ADAPT'14); 9th International Conference on High-Performance and Embedded Architectures and Compilers (HiPEAC'14), Vienna, Austria. January 20-22, 2014. *Best Paper Award*
Tom St. John, Benoit Meister, Andres Marquez, Joseph B. Manzano, Guang R. Gao, and Xiaoming Li.

"Position Paper: Locality-Driven Scheduling of Tasks for Data-Dependent Multithreading"
In Proceedings of Workshop on Multi-Threaded Architectures and Applications (MTAAP 2014), May 2014, Accepted.
Jaime Arteaga, Stephane Zuckerman, Elkin Garcia, and Guang R. Gao.

"Toward a Self-aware System for Exascale Architectures"
In Proceedings of Euro-Par 2013: Parallel Processing Workshops; the 1st Workshop on Runtime and Operating Systems for the Many-core Era (ROME 2013), Aachen, Germany. August 26th, 2013.
Aaron Landwehr, Stephane Zuckerman, and Guang R. Gao.

"Runtime Systems for Extreme Scale Platforms"
Ph.D Thesis, December 2013
Sanjay Chatterjee

"Isolation for Nested Task Parallelism"
The 29th International Conference on the Object-Oriented Programming, System, Languages and Application (OOPSLA), October 2013
Jisheng Zhao, Roberto Lublinerman, Zoran Budimlic, Swarat Chaudhuri, Vivek Sarkar

"Bounded memory scheduling of dynamic task graphs"
To appear in The 23rd International Conference on Parallel Architectures and Compilation Techniques (PACT 2014)
Dragos Sbirlea, Zoran Budimlic, Vivek Sarkar

"Expressing DOACROSS Loop Dependencies in OpenMP"
9th International Workshop on OpenMP (IWOMP), September 2013
Jun Shirako, Priya Unnikrishnan, Sanjay Chatterjee, Kelvin Li, Vivek Sarkar

"The Flexible Preconditions Model for Macro-Dataflow Execution"
The 3rd Data-Flow Execution Models for Extreme Scale Computing (DFM), September 2013
Dragoș Sbîrlea, Alina Sbîrlea, Kyle B. Wheeler, Vivek Sarkar

"A Tale of Three Runtimes."
To appear in arXiv.org
Nicolas Vasilache, Muthu Baskaran, Tom Henretty, Benoit Meister, M. Harper Langston, and Richard Lethin

Websites

- X-Stack wiki - https://xstackwiki.modelado.org, which is now the public central point of reference for the overall X-Stack program, including content from the X-Stack Kickoff Meeting in September 2012, the PI Meetings in March 2013 and May 2014, and each X-Stack participant's public website.
- Traleika-Glacier X-Stack's public website - https://xstackwiki.modelado.org/Traleika_Glacier
- Open Community Runtime - https://01.org/open-community-runtime
- HCLib: http://habanero-rice.github.io/hclib/
- CnC-OCR: https://github.com/habanero-rice/cnc-ocr
- Traleika Glacier X-Stack private collaboration wiki - https://xstack.exascale-tech.com/wiki/index.php/Main_Page, with over 150 members from DOE, universities, and private industry. Membership is open to DOE collaborators from FastForward and DesignForward communities, as well.

Networks or Collaborations Fostered

Intel, Rice, Reservoir Labs, ETI, and UIUC are closely collaborating in defining and implementing the OCR interfaces and technologies enabling the interoperability of OCR with the technologies developed at those institutions.

We actively participated in SC13 in November 2013, in which we presented "The Open Community Runtime (OCR) Framework for Exascale Systems" as part of the Birds of a Feather session, hosted a collaboration workstation at the Intel booth to promote the X-Stack vision, and presented "Event Driven Task Runtimes for Extreme Scale" in the Intel theatre.

We participated in the Runtime Systems Summit held April 2014 - https://xstackwiki.modelado.org/Runtime_Systems

We jumpstarted formation of the Modelado Foundation in April 2014 – http://www.modelado.org. Modelado Foundation's mission is to support and coordinate the open development of innovative computational models and simulations in engineering, design, and the sciences. Their goal is to build a robust, inclusive, sustainable ecosystem for advanced, large-scale, parallel computing efforts.

As described above, we are collaborating with LLNL (Jim Belak) to refactor LULESH and CoMD kernels and with john Bell (LBNL) to refactor AMR/miniGMG to our environment for validation, testing, and demonstration.

We attend all the X-Stack meetings, meet with the community, actively participate with the members of the co-design centers, and collaborate with open sharing of information and results.

Rice University has begun extensive collaboration with research groups outside of this project based on the accomplishments in this project:

1. Collaboration with LBNL on integrating OCR and HCLib with UPC and GasNet.
2. Discussion with ORNL on integrating OCR runtime with the UCCS communication runtime.
3. Collaboration with Texas Instruments on implementing OCR on top of the TI Hawking heterogeneous platform.

The codelet model proposed by the University of Delaware has been an inspiration for OCR, and was selected to be run on a dataflow-inspired architecture model in the European Community funded TERAFLUX project (http://teraflux.eu). Jack B. Dennis at MIT is basing part of his work on the fine-grain Fresh Breeze model on the Codelet model. In addition, internal collaboration within UD was fostered with other professors in the department (Xiaoming Li and Chengmo Yang).

Technologies, Techniques, and Other Products

Intel	Technology outlook, hardware guidance, system architecture Concurrent collection (CnC), partner with OCR development System software
ET International	Functional simulator
Reservoir Labs	R-Stream compiler (Proprietary Commercial Product) Algorithms, and select application kernels
Rice University	Habanero-C Open Community Runtime (OCR)
U of Illinois	Hierarchical Tiled Arrays Microarchitecture, System architecture
U of Delaware	Event Driven Task (Codelet) execution model System software for resource management Self-aware system which leverages ETI's energy model and implements a heat model on top of it
UC San Diego	Application performance and tuning
PNNL	Applications, co-design interface

Inventions/Patent Applications:

None at this time.

Cost Status and Unexpended Funds

Our current cost picture is shown in the table below. The project commenced 9/1/12. The project years run from 9/1 to 8/31 of the following year.
