AD_____

Award Number: DAMD17-02-2-0048

TITLE: Monitoring and Mining Data Streams

PRINCIPAL INVESTIGATOR: Stanley B. Zdonik, Ph.D.

CONTRACTING ORGANIZATION: Brown University
Providence, Rhode Island 02912

REPORT DATE: October 2004

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2004	Annual (15 Sep 2003 - 14 Sep 2004)

4. TITLE AND SUBTITLE
Monitoring and Mining Data Streams

5. FUNDING NUMBERS
DAMD17-02-2-0048

6. AUTHOR(S)
Stanley B. Zdonik, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Brown University
Providence, Rhode Island 02912

E-Mail: sbz@cs.brown.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)
This work is preliminary since funding for this piece was only received 6 months ago. We have so far managed to complete a revision of the Borealis code base making it much more usable for general applications. We have worked closely with personnel from USARIEM to identify special processing needs for PAN's in a WPSM setting and have redesigned major portions of the Borealis code base to reflect this. It takes a novel position with respect to dealing with failure in a sensor network.

14. SUBJECT TERMS
No subject terms provided.

15. NUMBER OF PAGES
6

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Cover.....1

SF 298.....2

Table of Contents.....3

Introduction.....4

Body.....4

Key Research Accomplishments.....5

Reportable Outcomes.....6

Conclusions.....6

References.....6

Appendices.....

INTRODUCTION

This work is preliminary since funding for this piece was only received 6 months ago. We have so far managed to complete a revision of the Borealis code base making it much more useable for general applications. We have worked closely with personnel from USARIEM to identify special processing needs for PAN's in a WPSM setting and have redesigned major portions of the Borealis code base to reflect this. It takes a novel position with respect to dealing with failure in a sensor network.

BODY

Distributed Stream Processing and Sensors

The Aurora stream processing engine was a very valuable exercise to gain understanding about the basic technical questions that must be addressed by any stream processing engine. These issues included memory management, tuple scheduling, and load control. Aurora was designed to run on a single server, and its primary optimization goal was low-latency processing. This set of assumptions was chosen because it gave us the best opportunity to study the fundamentals, and it was useful for a large class of monitoring applications.

We have been designing a new architecture for the Borealis stream processor that more closely matches the requirements of sensor processing. In particular, Borealis addresses distributed operation as well as optimization for power consumption and bandwidth.

In the case of distributed operation, we are addressing the issues of automatic load balancing and fault-tolerance. In this setting, load is imposed by the network of operators that is processing a set of input streams. Thus, load balancing consists of algorithms for moving operators from one computing node to another, while the system is running. Fault tolerance takes several forms. In its more classic form, it supports redundant computing elements that are synchronized in such a way that a primary node can "fail over" to a secondary node when a failure is detected. To date no one has adapted high-availability algorithms to operate efficiently in a streaming context. We have done that and have published a paper [HBR05] on the topic.

Classic high-availability is too strong a guarantee for many sensor-based environments. Instead, an approach that adapts to failures by trading accuracy or confidence for continued operation is more appropriate. For example, in a sensor-based environment, it is more reasonable to react to a failure by perhaps reducing the accuracy of the results. Of course, a good adaptive algorithm will minimize the loss in confidence by using its resources intelligently. We have written and published a paper [TBH05] on this topic jointly with folks from USARIEM.

We have been working with scientists at the Army Soldier Systems Labs in Natick, MA on a problem of sensor network data management. The goal is to find a way to fit the

WPSM problem into the Borealis framework. In order to do this, some extensions are required to the Borealis architecture.

In particular, we have come up with a way of capturing multiple processing models for a given physiological state. The system will select which processing model to use, based on the availability of inputs. It will also adjust sampling rates (for sensors that can do so) dynamically in order to place the confidence in an acceptable range.

Borealis Development Progress

We have put some effort into improving the infrastructure of the Borealis code base. The installation and build processes have been enhanced and documented to make them easier to learn and use. The source code directory tree has been restructured to make the code base extensible and to optionally build modules. I upgraded the code to work with the latest versions of external code packages and tools.

Scripts have been written to access the revision control system. They are easier to use than raw commands and are more reliable as they detect collisions that occur when multiple developers are working on the same module.

Borealis has been changed to remove many hardware dependencies. Portable schema types were introduced for use by Borealis applications. To make the system source code itself more portable, machine independent data types have been declared and code has been modified to use them.

A design and development effort is underway to provide a new programming interface for applications. The new XML-based interface will accommodate new features and will be much easier to use than the current interface. The goal is to provide the tools and documentation to enable end users to write Borealis applications with a minimum of effort. A distributed catalog is also in the design phase. It will replace the current central catalog so that it will scale up for multiple processors.

A website for developers has been created to disseminate information that on Borealis development issues and to exchange details about ongoing projects. The website is located at:

`http://www.cs.brown.edu/research/borealis/developer/`

KEY RESEARCH ACCOMPLISHMENTS

- A redesign of the Aurora stream processing engine (now called Borealis) to manage complex resources in a distributed environment.
- A new theory of how to do confidence-based resource management in a failure-prone environment.

- A simulator to assist USARIEM in understanding various parameterizations of their wireless PAN.

REPORTABLE OUTCOMES

- A simple, demo-able version of Borealis
- A well-received demo of Borealis at this year's SIGMOD, the most prestigious database conference.
- Helpful feedback to USARIEM.

CONCLUSIONS

So far, we have been very successful at producing a novel infrastructure for distributed stream processing. We are in the process of fitting it into a sensor-based environment at USARIEM. The system will be able manage power consumption and bandwidth utilization automatically. It will tradeoff accuracy (confidence) with the use of resources.

REFERENCES

[AAB05] "Design Issues for Second Generation Stream Processing Engines", (D. Abadi, Y. Ahmad, M. Balazinska, U. Cetintemel, M. Cherniack, J. Hwang, W. Lindner, A. Maskey, N. Tatbul, Y. Xing, S. Zdonik), *Proceedings of the Conference for Innovative Database Research (CIDR)*, Asilomar, CA, January, 2005.

[BBC04] "Retrospective on Aurora", (H. Balakrishnan, M. Balazinska, D. Carney, U. Cetintemel, M. Cherniack, C. Convey, E. Galvez, J. Salz, M. Stonebraker, N. Tatbul, R. Tibbetts, S. Zdonik), *VLDB Journal: Special Issue on Data Stream Processing*, 2004. *To appear.*

[HBR05] "High-Availability Algorithms for Distributed Stream Processing", (J. Hwang, M. Balazinska, A. Rasin, U. Cetintemel, M. Stonebraker, S. Zdonik), *Proceedings of the International Conference on Data Engineering* (ICDE), Tokyo, Japan, April, 2005.

[TBH05] "Confidence-based Data Management for Personal Area Sensor Networks", (N. Tatbul, M. Buller, R. Hoyt, S. Mullen, S. Zdonik), *International Workshop on Data Management for Sensor Networks (DMSN'04)*, Toronto, Canada, August 2004.

[XZH05] "Dynamic Load Distribution in the Borealis Stream Processor", (Y. Xing, S. Zdonik, and J. Hwang), *Proceedings of the International Conference on Data Engineering* (ICDE), Tokyo, Japan, April, 2005.