Emergent Behavior
Intelligent Software Agents and Cluster Computing (ISAACC)

Strength thru Collaboration

Explanations should be “as simple as possible, but no simpler.”

-A. Einstein

Mark T. Elmore  Software
Agent
Research

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY
Oak Ridge National Laboratory

- Established in 1943 for the World War II Manhattan Project.
- ORNL today pioneers the development of new energy sources, technologies, and materials.
- The advancement of knowledge in
  - Biological, Chemical,
  - Computational, Engineering,
  - Environmental, Physical, and Social Sciences.
- Budget: $1B+, 80% Department of Energy, 20% work for others.
- 3800 employees, 1500 scientists and engineers
- + “Visiting Scientists”
ORNL Is Committed to the Knowledge Discovery Agenda

- Entire Research Division Focused on Knowledge Discovery
- Appropriate Resources: HPC, Networking, MRF, JICS
- LDRD Initiative in Knowledge Discovery
- Programmatic efforts well-aligned with this science agenda
Our Focus on Knowledge Discovery

- Actionable insights from massive, dynamic, disparate data sources
- Ability to ask more complex questions and detect more complex processes using higher data resolution and innovative data representations

<table>
<thead>
<tr>
<th>Discovered Data</th>
<th>Dynamic Disparate Sensor Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental Data</td>
<td>Massive Simulated Data</td>
</tr>
<tr>
<td>Pattern Detection</td>
<td>Data Driven Simulations</td>
</tr>
<tr>
<td>Infer Physical Law</td>
<td>Hypotheses Generation</td>
</tr>
<tr>
<td>Model Development</td>
<td>Heightened Sense of Urgency</td>
</tr>
<tr>
<td>Model Validation</td>
<td>Actionable Knowledge</td>
</tr>
<tr>
<td>New Science</td>
<td>Decision Support</td>
</tr>
<tr>
<td>New Technology</td>
<td></td>
</tr>
</tbody>
</table>
CSED Knowledge Discovery Capabilities

1. RTKD Middleware and Applications
2. Data Models and Systems
3. Text Analysis
4. Disparate Data Anomaly Detection
5. Empirical Syndromic Surveillance
6. Discrete Event Simulations
7. Anticipatory Methods Research
8. Sensor Network Systems
9. Trust and Information Operations
10. Disruptive Information Systems
11. Agent-based Computing
   - Agent-based Simulation
   - High-performance Agents-on-Clusters
12. Population Dynamics
13. Geo-Spatial models
Why Cluster Computing?

- With the tsunami of data currently available, many current computing problems are tractable only by restricting the amount of data
- Incomplete or restricted answers come from incomplete or restricted data
- Cluster computing offers a relatively affordable platform
  - Multiply compute power using off-the-shelf components
- Leveraging a strong relationship with ORNL’s Cluster Computing research group
  - Well-designed architectures deliver future-thinking, future-capable designs
  - Increases usable production life of cluster system
Why ORNL Agents?

- High-Volume ✓ High-Speed ✓ Real-Time
- Scalable Parallel Processing Solutions
- Fault-tolerant and Recoverable Solutions
- For problems with *very large, distributed and diverse* datasets
  - Agent-based data fusion
  - Agent-based analysis
  - Across multiple data sources
- Top-notch ORNL scientists *collaborate* with top-notch scientists across the country
  - *Deliver effective* agent-based solutions
Why Agent-Based Cluster Computing?

- Having compute power is not enough
  - It is all about harnessing the available compute power

- Software Agents can manage and coordinate algorithms
  - Optimize their parallelization across the nodes of a Cluster Computer

- ORNL Software Agents have a proven track record
  - Providing a succinct, easily-understood mechanism to manage cluster computing parallelization

- Parallelization of Emergent Behavior Agents can bring increased capabilities
Meeting the Needs of the Tech Savvy Sponsor

- Sponsors’ needs for HPC
  - Have outgrown basic research
  - Sponsors now need functional prototypes and production-grade systems
- ORNL is in a *unique position*
  - Established capability for production-capable systems
    - Software Agent approaches to solving sponsor’s problems
  - Established relationship between
    - Cluster Computer Research
    - Software Agent management of that resource
- ORNL
  - Early recognition of this growing need
  - Established capabilities to meet this growing need
  - Has demonstrable examples
  - Record of building collaborative teams with sponsors
Emergent Behavior for Knowledge Discovery

• “The arising of novel and coherent structures, patterns and properties … in complex systems.
  1. “Radical novelty (features not previously observed in systems)
  2. “Coherence or correlation (meaning integrated wholes that maintain themselves over some period of time)
  3. “A global or macro "level" (i.e. there is some property of wholeness)
  4. “The product of a dynamical process (it evolves)
  5. “Is "ostensive" - it can be perceived.”

• Jeffery Goldstein, Founding Editor of Emergence
Emergent Behavior for Knowledge Discovery

- Swarm Intelligence and Software Agents are unique software engineering approaches that can provide tractable solutions to complex and difficult problems.
- While each individual agent is relatively simple and unsophisticated
  - The collective actions of the individual agents can exhibit complex and sophisticated behavior.
- The emergent behaviors provide a collective of distributed processes
  - With little or no centralized control
  - Across a variety of disparate hardware platforms.
Emergent Behavior for Knowledge Discovery

- Exhibit the ability to adjust to the dynamics of distributed computational environments,
  - Find optimal solutions based on the run-time resources available during the solution
  - Thus providing non-deterministic solutions that can exceed typical deterministic software engineering designs.

- With the globalization
  - Of data,
  - Of algorithms,
  - And of the hardware resources to bring them together,

The emergent behavior of software agents are a powerful approach to large and complex problems
What are Software Agents?
Object Oriented Programming

- Encapsulation: Objects tie processing to data
- The idea of [Agents] departs significantly from that of object-oriented (O-O) programming
  - O-O strongly suggests binding data and its processing together
- So, in O-O programming style
  - Every CD would come with its own player
  - And they are not supposed to be separated.
- As odd as this sounds
  - This the way we have built many software systems
~Hao He
Agents, on the other hand

- Separate processing and data
- And separate the hardware they run on
- In the CD analogy
  - CDs could play on multiple CD players
  - And you may take a CD and a CD player to a number of locations
- With Agents
  - Processing Agent
  - Meets Data Agent
  - At Agent Host with appropriate computing power
This Meeting Is An Agent Example Exhibiting Emergent Behavior

• We didn’t have to know ahead of time
  – Who specifically would show up
  – What the specific facilities would be
  – Or even what the specific topics would be
• And these can change throughout the meeting
• Yet we work together
  – Our Emerging Behavior generates the best solution given the capabilities present
Emergent Behavior Architectures
RTKD Conceptual Global Architecture

Distributed Capabilities

Distributed Connected Platforms

Distributed Regional Databases

KD Layer
- Fast Search
- Anomaly Detect
- Data Validation
- CDP
- HCI
- Deception Detect
- Dynamic Q&A
- Etc

Infrastructure Layer
- SOA
- HPC
- Hand Held
- Desktops
- Compute Clusters
- Etc

Data Layer
- DB
- DB
- DB
- DB
- DB
- DB
- DB
- DB
- DB

mte 2007
Agent-based Parallelization
Unparalleled Capability: Parallel Dynamic Text Analysis

• Breakthrough:
  – ORNL Software Agents Coordinating Distributed Collections

• Result:
  – Software Agents can now organize larger document sets by using multiple computers
    • Agents can cooperate across network distributed machines
    • Empowers Software Agents To Harness High-Performance Cluster Computing
    • High-Volume, High-Speed, Real-Time Solutions
Document De-duplication
US Pacific Command (PACOM)
DIA’s Pacific RSC

- **Agents-on-Cluster Research**
  - Red Oak / White Oak Twin Clusters
  - DIA Owned
  - ORNL designed, built, managed
  - 1.7 TFLOPS in small footprint

- **High-Speed De-duplication**
  - Adjustable threshold
  - 10,000 documents de-duped in under 3 minutes

- **Parallelization of PA-RSC “Differ” tool**
  - PA-RSC developed
  - ORNL ISAACC distribution for order-of-magnitude increase

**Red Oak / White Oak Twin Clusters**

- 4 Dell 2850s each with
  - 3.2 GHz Dual Processor
  - 2 GB Ram
  - 430 GB Disk

- 131 Dell 1850s each with
  - 3.2 GHz Dual Processor
  - 2 GB Ram
  - 73 GB Disk

**Total**
- 270 3.2 GHz Processors (Hyper-Threading available)
- 270 GB Memory
- 11.3 TB Disk
- 1.7 TFLOPS
Automatic Disambiguation Initiative
ADI Parallelization
ADI Solution

- The goal of the ADI effort is to develop software that will parallel the activities and logical decisions that humans make while performing the disambiguation effort.
- Since the total number of entities exceeds 2.9 million, it is impractical and impossible for human efforts alone to disambiguate the entirety of the database.
- Software Agents mimic the approach taken by humans
  - But at High Performance Computing speeds
Autonomous Data Retrieval (ADR)
ADR Parallelization
ADR Solution

• The goal of the ADR effort is to find OSINT information on disambiguated items

• Other U.S. Government sponsors with emerging requirements:
  – That could leverage the work performed on ADI to solve mission critical issues within their own organizations
  – That could leverage the future work performed on ADR to mine OSINT for probabilistic alignment with their entities

• Program accomplishments thus far show strong potential in other environments
Topic Monitoring Parallelization
Topic Monitoring

The Problem – Measuring Human Perception

• ORNL/Private Company team to monitor media sources
  – TV, websites, etc
  – For public perception and topics of interest

• PC was finding that their data ingest rate was suffering due to lack of scalability and resources
  – Too many sources to monitor, not enough resources to monitor them all
  – Too costly to develop an in-house custom distributed solution
Topic Monitoring

The Solution

- **Enhanced Capability**
  - ORNL used their agent frameworks to wrap PC components and distribute them on the cluster
  - Solution was able to leverage prior building blocks, tools, and experience
  - The cluster-based solution was 1500% faster than the single-node solution, and this number is the baseline

- **Coupled with ORNL Knowledge Discovery**
  - Integration of ORNL Knowledge Discovery collaborative tools
**System Overview**

**ORNL Cluster**
- Ingest Server Agents
- Database Federation
- Query Agents
- PIRANHA Agents

**SOFTWARE AGENTS**
- ~75 Million Indexed Records
- ~500,000 records ingested per day
- 150+ Distributed Databases

**Collectors**
- Mississippi Collectors
  - ABC WAPT
  - NBC WLBT
  - CBS WJTV
  - PBS WMPN
- National Collectors
  - CNN & Headline News
  - FOX News
  - MS NBC
- Other Collectors
  - Reuters
  - Associated Press
  - Google News
  - RSS Blog

**ORNL Software Agent Research Solutions**
- High-Volume
- High-Speed
- Real-Time
ORNL Infrastructure Monitoring
Parallelization

ORNL SOFTWARE AGENT RESEARCH SOLUTIONS
✓ High-Volume ✓ High-Speed ✓ Real-Time
Infrastructure Parallelization: Proof of Concept

The Problem – Too many computations, not enough computer cycles

- Networks can quickly become too large for a single computer to process
  - Algorithms such as “shortest path” may take a considerable amount of time to run on a large dataset
  - And that time may not be available

- A more scalable solution is needed
  - To process large networks
  - Or to run algorithms on many networks simultaneously

- Too costly to develop an in-house custom distributed solution
Proof of Concept

1. Started with the original Infrastructure software.
2. Designed and implemented a single agent running on a single host and verify.
3. Deployed multiple agents across multiple nodes of a cluster and verify the parallel-produced results.
Sample Results

- Best results to date: first available agent approach with 64 work nodes
  - Ran 131,072 route requests in
    - 8,145 seconds
    - 135.75 minutes
    - 2.2625 hours
    - Average of 16.09 requests per second.
  - FortiusOne tested on a single processor with route requests taking 4 seconds per route.
- Roughly an *64X improvement* with ORNL’s cluster and agent system
CSED Knowledge Discovery in Cyber Security

- A teaming of the Cyber Security Group and the Applied Software Engineering Group
Real Time Knowledge Discovery in Cyber-Security
Example: Spear-Phishing Defense

Phishing: -$n$, (1996) An attempt to fraudulently acquire sensitive information

Spear-phishing: -$n$, (~2005) A highly-targeted phishing attack

- Leveraging: A synergistic collaboration of three ORNL research groups
  - Applied Software Engineering Research group (ASER)
  - Cybersecurity and Information Infrastructure Research group (CSIIR)
  - Network and Cluster Computing group (NCC)

- Solution: A Software Agent Knowledge Discovery approach to provide multi-level defense
  - In isolation, each defense provides a strong response to the problem
  - But in unison, a powerful synergist effect for significant spear-phishing defense
  - Intelligent Software Agents and Clustering Computing (ISAACC)

Defense Triad
- Agent-enabled Knowledge-base (Expert System) approach
- Agent-based Clustering/Similarity/Anomaly approach
- And an Agent-based Enterprise Protection Architecture
Real Time Knowledge Discovery in Cyber-Security
Spear-Phishing Defense Triad

1. Agent-enabled Knowledge-base
   A. a.k.a. Expert System or Rules-Based Approach
   B. A series of rules are captured in machine readable format (XML or derivative such as RDF)

2. Agent-based Clustering/Similarity/Anomaly Approach
   a. ORNL patented Intelligent Software Agents and Clustering Computing (ISAACC) approach
      • High-Performance Computing for complex algorithms
      • Multi-level comparisons at message, paragraph, sentence, and/or phrase level
   b. Messages compared with archive of phishing messages
   c. Messages with strong similarity to archive are then be appropriately handled.
      • Warning header attached, security notified to bring a human in the loop, etc

3. Agent-based Enterprise Protection Architecture
   a. Agents at each protected site (HQ, labs, etc) would automatically, and in real-time, share information about potential attacks
   b. SOA connectivity for flexibility, emergent behavior
   c. Information on an attack focused on one site
      • Can immediately trigger a proactive response at other sites
Summary / Discussion

- Agent-based solutions parallel the emergent behavior of human-based solutions
- ORNL is a leader in bringing agent-based solutions to bear
- Using computing resources
  - Where they are found
  - In a manner that utilizes their capabilities

Mark T. Elmore
ElmoreMT@ornl.gov
865-241-6372