Fast Primer Search with DUP

Christian Grothoff
christian@grothoff.org

Technische Universität München
Overview

• Research Area

• Easy Distributed Stream Processing with DUP

• Case Study: Fast Primer Search
  – Biological Questions
  – Primer Search Parallelization with DUP
  – Mapping Primers to Species with the BGRT

• Other Research Results

1“An algorithm for the comprehensive search of oligonucleotide signatures based on phylogenetic trees”, joint work with K. Bader, W. Ludwig and H. Meier
Problem Domains

- Systems
- Programming Languages
- Software Engineering
- Secure Networking
- Privacy

Runabout
\[
\text{HashMap\{Class,Code\}} \rightarrow \text{map}
\]
\text{visitAppropriate}(\text{Object}) : \text{void}

RunaboutSum
\[
\text{sum : int}
\]
\text{visit}(\text{A0}) : \text{void}
\text{visit}(\text{A1}) : \text{void}
\text{visit}(\text{A2}) : \text{void}

GenCodeXX
\[
\text{accept}(\text{Object}) : \text{void}
\]
The Problem:
Developing Parallel Stream Applications

- Most developers (only) know how to write sequential code
- Parallel programming is error-prone (data races, deadlocks)
- High-performance parallel programming is really hard
- With GPUs for $4,000, we could have 2,600 cores...

⇒ Developers more expensive than hardware
X10 vs. the DUP System

X10

10x faster, 10x as productive in 10 years for BlueGene

DUP

$\frac{1}{2}$ the speed, 10x as productive in 10 months for POSIX

Available at http://dupsystem.org/
Agenda

• Research Area

• Easy Distributed Stream Processing with DUP

• Case Study: Fast Primer Search
  – Biological Questions
  – Primer Search Parallelization with DUP
  – Mapping Primers to Species with the BGRT

• Other Research Results
The DUP System

- High-level DUP language
- DUP compiler
- Low-level DUP language
- dupd
- dupd
- dupd
- dup
- grep
- grep
- grep
- in.txt
- fanout
- TCP
- TCP
- TCP
- TCP
- faninany
- out.txt
DUP Applications

- Fast primer search
- High-throughput, customizable video-conferencing
- Parallel and distributed discrete event simulation framework
Agenda

• Research Area

• Easy Distributed Stream Processing with DUP

• Case Study: Fast Primer Search
  – Biological Questions
  – Primer Search Parallelization with DUP
  – Mapping Primers to Species with the BGRT

• Other Research Results
Biological Questions

• Which are the most specific OSSs for a species?
• Which are the most specific OSSs for a subtree in the phylogenetic tree?
Input: Phylogenetic Tree

```
I
  IIa
  IIb
    IIIa
    IIIb
II
  I
  IIa
I
```
Parallel Mapping of Primers to Species

s @opt1:88[0<in.txt,1|p1:0,3|p2:0] $ fanout;
p1@opt1:88[1|pe:0] $ arb_probe_dup;
p2@opt2:88[1|pe:3] $ arb_probe_dup;
pe@opt2:88[1>out.txt] $ gather;

![Graph showing speed-up vs. number of nodes for different mapping strategies]
Intermediate Result: Species and OSS
Agenda

• Research Area

• Easy Distributed Stream Processing with DUP

• Case Study: Fast Primer Search
  – Biological Questions
  – Primer Search Parallelization with DUP
  – Mapping Primers to Species with the BGRT

• Other Research Results

fsnsg
• Which are the most specific OSSs for a species?

• Which are the most specific OSSs for a subtree in the phylogenetic tree?

• Sequence information may contain errors; allow up to $k$ out-group hits!

• Perfect OSS may not exist even with out-group hits; maximize number of in-group hits
BGRT Creation (1/5)

∅ \xrightarrow{1,2 : A} ∅

\text{fsnsg}
BGRT Creation (2/5)
BGRT Creation (3/5)

∅

1,2 : A
2,3 : B

1,3 : C

∅

1

2 : A
3 : C
2,3 : B

fsnsf
BGRT Creation (4/5)

\[
\varnothing 
\begin{array}{c}
\mid \\
1 \quad 2 : A \\
\mid 3 : C \\
\mid 2,3 : B \\
\end{array}
\quad \leftrightarrow \quad \begin{array}{c}
\varnothing \\
\mid \\
1 \quad 2 : A \\
\mid 3 : D \\
\mid 3 : C \\
\mid 2,3 : B \\
\end{array}
\]
BGRT Creation (5/5)

∅

- 1 : E
  - 2 : A
    - 3 : D
    - 3 : C
  - 2,3 : B
  - 5 : G
    - 4 : F
    - 6 : H
Iterate-And-Bound

```
fsnsg
```

```

1 : E
2 : A
3 : D
3 : C
2,3 : B
5 : G
4 : F
6 : H
```
## Final Result

<table>
<thead>
<tr>
<th>Mismatches</th>
<th>I</th>
<th>IIa</th>
<th>1</th>
<th>2</th>
<th>IIb</th>
<th>IIIa</th>
<th>3</th>
<th>4</th>
<th>IIIb</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3:D,G</td>
<td>2:A</td>
<td>1:E</td>
<td>%</td>
<td>2:H</td>
<td>1:F</td>
<td>%</td>
<td>1:F</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>2</td>
<td>%</td>
<td>1:G*</td>
<td>1:D+</td>
<td>1:D,G+</td>
<td>1:D*</td>
<td>1:D,G+</td>
<td>1:D+</td>
<td>%</td>
<td>1:G+</td>
<td>1:G</td>
<td>%</td>
</tr>
<tr>
<td>3</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>0:D*</td>
<td>%</td>
</tr>
</tbody>
</table>

“+” Should probably not be computed (mismatches >, matches =)

“∗” Even more useless (mismatches >, matches <)

\[ \text{fsnsg} \]
Agenda

• Research Area

• Easy Distributed Stream Processing with DUP

• Case Study: Fast Primer Search
  – Biological Questions
  – Primer Search Parallelization with DUP
  – Mapping Primers to Species with the BGRT

• Other Research Results
Other Current Work

- Successful attacks on various “secure” P2P networks (Tor, Freenet, Tahoe LAFS, ...)
- Randomized Resilient Routing in Restricted Route Topologies
- Autonomous NAT traversal
- Automatic Restart Management: RAS for GNU/Linux
- Parallelizing Protein Analysis (with Rost Lab)
An Attack on Tor (USENIX Security 2009)
Future Work

- Resource allocation for DUP
- Parallelize more applications with DUP
- Aspect-oriented coordination language for DUP
- Migrating to IPv6 using secure P2P VPN over GNUnet
- Memory fragmentation analysis (Mallice)
Questions
The GNUnet Framework

- **Transport**
  - TCP, UDP, HTTP, ...

- **Routing**
  - DV
  - GAP
  - DHT

- **Encryption**

- **Authoring**

- **Datastore**
  - MySQL
  - Postgres
  - sqlite

- **ARM**

- **Testing**

- **fsnsg**
GNUnet’s Automatic Restart Manager

- User requires services
- Dependency Learning and Management
- Process
  - crashes
    - ARM
    - restarts impacted processes
    - ARM
    - investigates bugs
    - reports crash
- User
  - changes configuration
- Developer
  - monitors configuration changes
  - ARM
  - starts required services
  - ARM
  - dependency
  - uses dependency
- Bug Tacking System
  - reports crash
  - analyzes crash, adds instrumentation, restarts service
  - User
  - reports crash
  - User
  - User
Secure Routing (1/3)

Base Restricted Route Topology

A
B
C
D
E
F
G
H

A - B - C - D - E - F - G - H - A
Secure Routing (2/3)

Distance Vector Added Links

A - B - C - D - E - F - G - H
Secure Routing (3/3)

Resulting DHT Topology

```
000
010
100
111
001
110
101
011
```
Teaching

• Programming Languages
• Mainframe Administration
• Computer Networking
• Peer-to-Peer Networking
Group Infrastructure

Software:
• Drupal / Mantis
• Doxygen / Subversion
• Buildbot / lcov
• clang (LLVM)
• Coverity Prevent

Hardware:
• Server (i7)
• Sheeva Plug (ARM)
• iMac (PowerPC)
• GNU/Linux VMs (AMD)
• PMP-capable NAT router

Furthermore, we have access to a wide range of networking equipment (via Lehrstuhl) and HPC facilities at the LRZ.