Boosting Productivity with Advanced User Services

Raffaele Montuoro
Advanced User Services: Mission

• Enhance and support computational sciences within Texas A&M University
Advanced User Services: Details

• Performance Analysis of computer codes
• Code Optimization
• Code Parallelization: OpenMP & MPI
• Consulting for code development on SC systems
• Code Porting on the SC systems
• Tune up of common scientific applications
• Benchmarking
• Design and configuration of small computer clusters
Advanced User Services: Young, but Tall

Cumulative Number of Projects

months

2008 | 2009
---|---
5  | 16
13 | 13
16 | 16

Cumulative Number of Projects
Advanced User Services: A Closer Look

• SODA: A Simple Ocean Data Assimilation Model
  Dr. Benjamin Giese, Dept. of Oceanography

• MST: Material Simulation Tool
  Dr. Tahir Cagin, Dept. of Chemical Engineering

• Illumina Genome Analysis Pipeline
  Dr. James Sacchettini, Dept. of Biochemistry and Biophysics
SODA: Simple Ocean Data Assimilation

Original Code: FORTRAN/OpenMP

- CPU: 16
  - Original: 1.6x
  - OpenMP: 1.3x

- CPU: 32
  - OpenMP: 2.0x
  - OpenMP+MPI: 1.6x

- CPU: 64
  - MPI: 2.4x

Case: 2004-01-20

B. Giese

Ping Luo
Xiandong Meng
SC staff

1989-2009

Advanced User Services

6
Material Simulation Tool

Original Code: C++, serial

Case: 101010

- Original (serial)
- Optimized (serial)
- OpenMP

CPU

8

16

Speedup

1

3.1x

18.2x

23.3x

Original Code:

CPU Speedup

Case: 101010

Ping Luo

SC staff
Whole genome sequencing can be used to define the mechanism of drug action and resistance.
How Solexa sequencing works

Single-ended sequencing:
1) Fragment gDNA select 200-300 base fragments
2) Spread and attach fragments to a lane on a chip, amplify
3) Press go
4) Just align your short reads, allowing gaps/mismatches, against a reference genome, and look for snps and indels
What can you do with all that sequence?

Sequence an entire genome of Mtb — 4M bases in the Mtb genome

Therefore with 20Mx36 base reads you get 100% coverage with 200-fold redundancy- (depth of coverage) per genome; 7 genomes per chip

Or you can add a tag to each of 4 genomes and run them on a single lane -50 fold depth of cover per genome- 28 genomes per chip

Parallel Genome Analysis Pipeline

J. Sacchettini
T. Ioerger
**Ioerger Sequencing Protocol**

1. **DNA** → **Solexa** → file with millions of 36-bp reads

2. **align reads** against reference genome

3. Identify **putative SNPs** where majority base differs from expected base

4. Build **local contigs** in surrounding ~200 bp and align against reference genome to identify true SNPs vs. indels

5. Edit genome sequence and realign reads

**Parallel Genome Analysis Pipeline**

*J. Sacchettini T. Ioerger*
• Solexa’s Genome Analysis Pipeline is a customizable analysis engine capable of taking the raw image data generated by the Genome Analyzer and producing intensity scores, base calls, and quality metrics, and quality scored alignments

• Based on Makefile

• Scales up to 8 shared-memory tasks (gmake -j 8)

• Typical problem size: 8 lanes x 36 cycles x 4 bases x 100 images/base/cycle = **115,200** images to be processed
Parallel Genome Analysis Pipeline

ANALYSIS none, 79 cycles, **252,800 images**

<table>
<thead>
<tr>
<th>CPUs</th>
<th>Elapsed time (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>26, 19.95</td>
</tr>
<tr>
<td>32</td>
<td>12.56, 2.1x</td>
</tr>
<tr>
<td>64</td>
<td>7.42, 3.5x</td>
</tr>
<tr>
<td>128</td>
<td>5.03, 5.2x</td>
</tr>
<tr>
<td>256</td>
<td>3.32, 7.8x</td>
</tr>
</tbody>
</table>

- Solexa
- SC parallel

Raffaele Montuoro
SC staff
Parallel Genome Analysis Pipeline

ANALYSIS none, 79 cycles, **252,800 images**
Q: How to apply?

A: E-mail the Supercomputing Help Desk:

help@sc.tamu.edu