

## Big Computation and Big Data

<http://www.eidos.ic.i.u-tokyo.ac.jp/>

### 1 Introduction

Computing has become an indispensable tool in virtually all fields in science and engineering (e.g., for understanding the mechanism of biomolecules, predicting climate changes, designing new materials, understanding galaxy formation, etc.). They have been putting ever increasing demands on computing devices, and have been fulfilled only by large-scale parallel computers. For a familiar example, the IBM Watson computer, which in 2011 defeated the world human champion in Jeopardy!, a popular quiz show in US, used about 3,000 computers to generate and prune candidate answers. Another example is a search engine like Google, which utilizes parallel computers to collect a large amount of data from the web and build their indices. Fastest computers in the world as of 2012 connect several thousands to more than hundreds of thousands of computer nodes.

### 2 Theme

Our central themes are software systems for helping scientists/engineers perform high-performance parallel computation easily; the goal is to achieve both “programmability” and “performance.” Specific topics include:

- High level parallel programming languages for next-generation massively parallel computers
- Parallel file systems and databases for processing big data
- Shell/scripting environments for almost programming-free parallel processing

By designing and implementing such systems, we aim at publishing software widely used by other researchers/engineers.

CPU clock speed have plateaued in recent years and thus, serial programs can no longer enjoy “automatic” speedup (speedup without modifications). The parallelism exposed to the programmer will only increase in coming years; only by utilizing them effectively can software accelerate. In this sense, system software researches in near future is almost necessarily researches about parallel processing in one way or another.

### 3 Some Ongoing Projects

#### **MassiveThreads: Super Lightweight Thread Library:**

A thread library that can spawn threads two orders of magnitude faster than native operating system threads. Based on this technology, we are working on high-level

parallel programming languages that target a range of machines from multi-core computers to massively parallel processors (<http://code.google.com/p/massivethreads/>).

**ParaLite Parallel Database:** A parallel database that can easily parallelize workflow applications consisting of existing commands, using SQL syntax. We are building a large scale data processing system based on ParaLite.

**GXP Parallel Shell/Scripting Tool:** A parallel processing tool that enables parallel processing built on existing commands, no matter where you are and where your resources are (<http://www.logos.t.u-tokyo.ac.jp/gxp/>).

I recommend everybody in the lab to publish their research results not only as papers but also as free software others can use.

### 4 Computing Environments

- InTrigger ([www.intrigger.jp](http://www.intrigger.jp)) is an environment that connects 1,800 CPU cores across 17 sites, providing an environment for large scale distributed computation. Our group played a leading role in designing, building, and operating this environment.
- We are using cutting-edge supercomputers installed in University of Tokyo and Tokyo Institute of Technology for our research.
- We have large multi-core multiprocessor machines (with 24/32 cores).

They form an ideal environment for enjoying power and joy of parallel computing.

### 5 Whom do we like to join us and whom do we like to educate?

You may be afraid that you must be good at programming and computer systems already? Well it may be true to some extent in this field, but they are only a small part of what you will/should learn in your life anyways. So, what’s far more important is that you are motivated to learn by yourself, and that you keep thinking with your own brain or discussing with your colleagues, not trying to find “answers” with Google. We like such self-motivated students to join us and learn together with us.