About AIST
— Integration for Innovation —

The National Institute of Advanced Industrial Science and Technology (AIST) is one of the largest public research organizations in Japan, focused on “bridging” the gap between innovative technological seeds and commercialization. AIST is organized into 5 departments and 2 centers, bringing together core technologies that leverage on a strong multidisciplinary research environment.

The Information Technology Research Institute (ITRI) and The Artificial Intelligence Research Center (AIRC) are the two major groups in the Department of Information Technology and Human Factors of AIST.

ITRI research areas:
- Computer architecture and system software for next-generation data centers to achieve fast and efficient big data processing
- Cyber physical software engineering and advanced cryptosystems that facilitate secure and dependable cyber physical systems
- “Creativity enabling technology” that encourages the creation and utilization of valuable content

AIRC research areas:
- Development of flexible and robust computation methods for AI, which emulate various aspects of computation in the human brain
- Integration of machine learning inferences based on huge data with inferences based on knowledge and text that human society has accumulated
Flow-centric Computing

Highly efficient cloud-based machine learning system architecture and software toward the post-Moore era

We propose Flow Centric Heterogeneous Architecture (FCHA), in which heterogeneous computing engines are connected by a high bandwidth communication network and data are moved between engines aggressively so as to process the data as efficiently as possible. FCHA will be one of the ways to improve computing performance even after the end of Moore’s law. We recently initiated a project to develop a cloud-based machine learning system based on the FCHA. Flow in Cloud is a pool of multiple kinds of processing engines such as FPGA and GPU connected by a circuit switch network. Based on requirements from jobs, such engines are combined and connected together and provided to users by a special operating system called Flow OS.

Flow-centric Heterogeneous Architecture in the Post-Moore Era

Balance between computation and communication will change in the post-Moore era: We cannot expect significant computing performance improvement of a single LSI when Moore’s law comes to its end. On the other hand, communication performance in a system can be improved by exploiting optical communication technologies, including polarization and wavelength division multiplexing, for wide area networks. These technologies are expected to increase the bandwidth to 10s of Tbps by 2030.

Centralized generator of wavelengths (wavelength bank: WB), and silicon photonics technology will realize low power and low cost optical communication system.

Heterogeneous special purpose processing engines: When Moore’s law ends, performance improvements have to be achieved through innovations in architecture and algorithm. Special purpose hardware will finally outperform general purpose computers. By combining different kinds of special purpose hardware and using specific types in specific stages of processing jobs, performance improvement can be achieved.

Data affinity to function affinity: Data moving cost will become relatively low, since 10s of Tbps is equivalent to memory bandwidth. By combining task specific processors in a pipeline manner, instead of using general purpose CPUs with large memory for all processing jobs, we can optimize the processing by matching data type to the function of specific processors.

Flow OS

Flow OS is a datacenter-wide operating system for managing and controlling the whole Flow in Cloud system. It provides a “slice” to each user according to job requirements. A slice is a set of connected computing nodes tailored for the job according to the flow of data. It also optimizes the total resource utilization by using hierarchical scheduling, distributed monitoring, and AI-based policy management. To properly provision an application execution environment for each slice, Bare Metal Container technology is employed (see page 9).

 Highly Efficient Cloud-based Machine Learning Architecture

Proof-of-Concept demonstration for FCHA: Flow in Cloud is a pool of different kinds of computing nodes with special purpose engines such as GPU and FPGA. Switch nodes provide a simple TDM (Time Division Multiplexing) circuit switch network. The controller is outside of the data flow; the data plane and the control plane are separated.

NEDO IoT Project

This project is supported in part by the Crosscutting Technology Development Project to Promote IoT, under NEDO (New Energy and Industrial Technology Development Organization), Japan. Through tight collaboration with academic and industrial participants, we will carry out R&D for an IoT immersed society.
Artificial Intelligence Research Center (AIRC)

It is essential to have a close relationship between application research and fundamental research for Artificial Intelligence (AI). In this relationship, the use of cutting-edge research in real-world problems brings about new, innovative technologies for society. The goal of AIRC is to become a research hub where both established and young researchers work together on joint projects, and where industries and academia can collaborate to achieve the same goal.

Goal of AIRC

AIRC is aiming at AI embedded in the real-world; i.e., AI that collaborates with human beings to solve difficult social problems.

The applications are categorized into three areas, namely:
- AI for Human Life / Services
- AI for Science / Engineering
- AI for Manufacturing

We are engaged in the following efforts:

(1) Target-oriented Fundamental Research
   - Brain-inspired AI and data-knowledge integration AI
   - Advanced machine learning / probabilistic modeling

(2) AI Framework Research / Development
   - Establish a framework for AI
   - Implement various elemental technologies, such as data collection, recognition, and prediction as pluggable 'core modules'.

(3) Standard Task / Benchmark Development
   - Develop standard tasks such as human action modeling
   - Develop Benchmark datasets

AI Bridging Cloud Infrastructure (ABCI)

ABCI is an open innovation platform with computing resources of more than hundred petaflops for world-class AI research and development. Through industry and academia collaboration, algorithms, big data, and computing power will be leveraged in a single common public platform. ABCI will rapidly accelerate the deployment of AI into real businesses and society. ABCI will be introduced in 2017-2018 at Kashiwa Campus, University of Tokyo by AIST sponsored by METI (Ministry of Economy, Trade and Industry) in Japan.

AI Infrastructure

Convergence of algorithm, big data, and computer power is a key issue for the success of AI research and development. ABCI employs highly integrated designs of modern HPC and big data technologies for the underlying hardware and software architecture. These designs accelerate workloads of various AI, machine learning, and deep learning applications.

- Extreme computing power with more than hundred petaflops tailored for AI, machine learning (ML) and deep learning (DL).
- Big data and HPC integrated in modern design for advanced data analytics, scientific simulation, etc.
- Ultra high bandwidth and low latency in memory, network and storage for accelerating various AI workloads.

Bridging Infrastructure

ABCI is an open innovation platform that facilitates industry and academia collaboration for AI research and development. By converging various AI companies in Japan, such as big data holders, ML/DL framework developers, and AI service providers, ABCI encourages technology transfers that bridge AI researches with real business and society. ABCI strongly accelerates the deployment of AI technology, including AI common technologies to startups and AI cloud design to IDC vendors.

- Common public platform for AI applications, services, and infrastructure designs, aiming technology transfer through industry and academia collaboration
- Open hardware and software architecture with accelerator support based on commodity devices

ABCI open innovation platform

- Multi-peta-byte-class sharable big data storage for AI research and development collaboration
- Cloud Infrastructure

To deploy advanced AI cloud design to IDC vendors, construction and operation cost must be reduced. ABCI aggressively employs novel but commodity components from the ground-up by leveraging next-generation supercomputer architecture and facility designs.

- Ultra-dense IDC design from ground-up on lightweight 'warehouse' building with substantial tolerance
- Wide-ranging Big Data and HPC standard software stacks for cloud ecosystem
- Extreme green with ambient warm cooling, large Li-ion battery storage, high-efficiency power supplies, etc.
- Advanced cloud-based operation including dynamic deployment, container-based virtualized provisioning, multitenant partitioning, automatic failure recovery, etc.

http://www.airc.aist.go.jp
Development of HPC Applications

Extensible & Scalable TCAD Simulator / Middleware for Fault-Resilient Programming

As a member of the multidisciplinary organization, we collaborate with other research departments of AIST to develop practical HPC applications. Among them, ImpulseTCAD is a novel Technology CAD simulator utilized for designing the next generation semiconductor devices. We also study on programming middleware that benefits the development of HPC applications. Impulse is such a middleware that helps implementation of scalable and fault resilient applications.

ImpulseTCAD

Technology CAD (TCAD) is a semiconductor device simulator, which is used as an industrial tool to design devices. As patterns of the semiconductor devices become finer, crosstalk between transistors becomes more significant. To predict characteristics of such devices, multiple transistors have to be simulated in an unified manner. In the mean time, several new materials are introduced to overcome physical limitation of the current silicon processing, requiring to develop and implement novel physical models in the simulator.

ImpulseTCAD is our project to develop an extensible and novel physical models in the simulator. To solve a set of semiconductor ImpulseTCAD is our project to develop an extensible and novel physical models in the simulator.

Nonlinear FVM is a non-linear FVM solver, which employs an automatic differentiation framework for extensibility. With Nonlinear FVM, users can concentrate on describing physical models, and all derivation code necessary for the Newton method are automatically generated.

ImpulseMesh is an unstructured mesh library, that provides involved geometric properties of 3D cells in implementing the vertex-centered FVM. It also offers access to boundaries between spacial regions, allowing flexible setup of boundary conditions.

ImpulseELAI is an elastic linear algebra interface. It offers an abstract and flexible constructor of sparse structures, as well as several interfaces to scalable sparse linear solvers.

Thanks to the easy extensibility, physical models of exotic materials as ferroelectrics are implemented into ImpulseTCAD, which is used to study the NC-FET devices.

Falnax

For the upcoming exascale computers consist of a huge number of components, and failures among them are no longer a rare event. Because of the unstable environment, exascale applications are required to be not only scalable, but also fault resilient. Plain MPI applications, however, will be terminated by a single failure, because data and tasks are statically assigned to ranks that cannot be recovered if a rank fails. To overcome failures and sustain calculations, complicated exception handling is necessary.

Falnax is our MPI middleware for the fault resilience, built on top of ULFM-MPI. In Falnax applications, most of the computational resources are delegated to a Falnax runtime, and application logic is described outside of the runtime as a workflow of tasks. The runtime is split into a data store (DS) and a set of workers, and failures among them are handled by the middleware. The tasks are implemented as MPI routines programmed to be portable among workers, and are processed on the workers. When a worker fails, the running task is terminated and rescheduled on other workers. Input/output data of tasks are stored in DS, where stored data are implicitly replicated to prevent data loss due to failures. In short, tasks are considered as local rollback units in Falnax, liberating users from complicated exception handling for the fault resilience.

Falnax is used in OpenFMO, an ab initio quantum chemistry program. In OpenFMO, DS is also utilized to exchange data directly among tasks.

Non-volatile Memory

RAMinate: Hypervisor-based Virtualization for Hybrid Memory Systems

STT-MRAM will achieve larger capacity and comparable read/write performance, but incur orders of magnitude greater write energy than DRAM. To achieve large capacity as well as energy-efficiency, it is necessary to use both DRAM and STT-MRAM for the main memory of a computer. We propose a hypervisor-based hybrid memory mechanism, RAMinate, that reduces write traffic to STT-MRAM by optimizing page locations between DRAM and STT-MRAM. It does not require any special program at the operating system level nor any design changes of the current memory controller at the hardware level.

Main Memory with DRAM and NVM

According to the technology roadmap forecasted by the semiconductor industry, Spin Transfer Torque Magnetoresistive RAM (STT-MRAM) will achieve the same level of read/write latency as DRAM around 2016, as shown in Table 1. In contrast to the reliability problems of Phase Change Memory (PCM), STT-MRAM can rewrite memory cells without any practical degradation. STT-MRAM has the potential of serving as the main memory of computer systems like DRAM does today.

However, one disadvantage of STT-MRAM is that the write energy will be 10^3 times larger than that of DRAM. Just replacing all DRAM with STT-MRAM will likely increase the energy consumption of a memory subsystem. To address this problem, a hybrid memory system that combines both large STT-MRAM modules and small DRAM modules for the main memory of a computer is necessary.

Although prior studies discussed the designs for hybrid memory systems, software-based mechanisms need to update operating systems or application programs, and hardware-based mechanisms need to modify the memory subsystem of computers. Both approaches are not transparent to existing software and hardware components, which will incur huge migration costs to take advantage of emerging NVM technologies.

<table>
<thead>
<tr>
<th>STT-MRAM and DRAM</th>
<th>2013</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Time (ns) DRAM</td>
<td>&lt;10</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Write/Erase Time (ns) DRAM</td>
<td>&lt;10</td>
<td>&lt;10</td>
</tr>
<tr>
<td>STT-MRAM 35</td>
<td>&lt;10</td>
<td>&lt;10</td>
</tr>
<tr>
<td>STT-MRAM 35</td>
<td>&lt;1</td>
<td></td>
</tr>
<tr>
<td>Write Energy (J/bit) DRAM</td>
<td>4E-15</td>
<td>2E-15</td>
</tr>
<tr>
<td>STT-MRAM 2.5E-12</td>
<td>1.5E-13</td>
<td></td>
</tr>
</tbody>
</table>

Hyervisor for Hybrid Main Memory

RAMinate is a hypervisor-based hybrid memory mechanism that optimizes page locations between DRAM and MRAM in a manner fully transparent to guest operating systems. It periodically monitors memory access of a virtual machine and dynamically determines write-intensive guest memory pages, and dynamically updates page mapping between guest and physical memory pages.

We developed a prototype of the proposed system by extending Qemu/KVM and conducted experiments with application benchmarks.

We confirmed that our page replacement mechanism successfully worked for unmodified operating systems and dynamically diverted memory write traffic to DRAM. Our system successfully reduced write traffic to STT-MRAM by approximately 70% for tested workloads, which results in a 50% reduction in energy consumption in comparison to a DRAM-only system.

This work was published in the 7th ACM Symposium on Cloud Computing (SoCC2016). Our paper is available online. http://doi.acm.org/10.1145/2987550.2987570

This work is partially supported by the IMPULSE project of AIST and JSPS Grant KAKENHI 16K00115.

Page remapping mechanism of RAMinate.

We confirmed that our page replacement mechanism successfully worked for unmodified operating systems and dynamically diverted memory write traffic to DRAM. Our system successfully reduced write traffic to STT-MRAM by approximately 70% for tested workloads, which results in a 50% reduction in energy consumption in comparison to a DRAM-only system.

This work was published in the 7th ACM Symposium on Cloud Computing (SoCC2016). Our paper is available online. http://doi.acm.org/10.1145/2987550.2987570

This work is partially supported by the IMPULSE project of AIST and JSPS Grant KAKENHI 16K00115.
Bare-Metal Container offers an environment to run a container image of Docker with a suitable Linux kernel on a remote physical machine. The Docker image is allowed to change the kernel and its settings. As a result, the application extracts the full performance of the physical machine. This mechanism allows changes to both the kernel and machine for optimum performance for a given application.

**Background**

As the popularity of Docker shows, container technology is becoming widely used because it enables easy customization and can be shared on hub servers. It is also popular for HPC applications because it offers quick execution. However, container technology does not allow optimizing the kernel for an application and cannot extract full performance on a target machine when considering power consumption. Therefore, container technology is system-centric architecture. The kernel for the container is not easily updated on the server, and the server wastes power when no application is running.

**Bare-Metal Container**

BMC: Bare-Metal Container solves these problems using remote machine management mechanisms (WakeUpOnLAN, Intel AMT, and IPMI), a network bootloader, and power measurement. BMC offers a suitable kernel for a Docker’s Intel AMT, and IPMI), a network bootloader, and power management. BMC is an application-centric architecture, while the traditional style is server-centric architecture. Furthermore, BMC is designed to be used on LAN and the Internet. It has two modes to obtain a root file system: NFS mode for secure LAN and RAMFS mode for insecure Internet. Users can select the mode depending on their environment.

Although BMC requires the overhead of kernel booting, HPC applications can improve performance, which surpasses the overhead. We have already evaluated the kernel optimizations for CPU (Hyper Threading), memory (Transparent Huge Pages), and network (Receive Flow Steering) were applied to an application on a number of machines, from low power CPU (Celeron) to high speed CPU (Xeon). The results showed affinities of the kernel optimization for applications and the improved performance could surpass the overhead caused by BMC.

On BMC, the power consumption for the application is measured, and the information is used for optimization. The machine is also shut down when an application is terminated. It means the energy is used for an application only. Therefore, BMC is an application-centric architecture, while the traditional style is server-centric architecture.

**Comparison between traditional container and BMC**

- **Traditional Style (Ex: container)**
  - Invoke app.
  - Select a physical machine
  - Boot the kernel & app
  - Invoke app.
  - Select a kernel

- **BMC**
  - Select a physical machine
  - Boot the kernel & app

<table>
<thead>
<tr>
<th>Server Centric Architecture</th>
<th>Application Centric Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power always up</td>
<td>Power frequently up/down</td>
</tr>
</tbody>
</table>

**Natural Disasters: Data-intensive Events**

A disaster is a dynamic event that progresses linearly through time and generates big data sets from a variety of sources. These data need to be acquired, aggregated, and analyzed in order to provide useful information to professionals responding to the disaster. With this in mind, the sharing of information is a critical component in the decision making process along the disaster event timeline. Different organizations and authorities may be geographically distributed, either because they are in physically distinct locations to begin with or because the affected disaster area has separated them. Therefore, these organizations must be able to interact with each other through the sharing and analysis of big data sets, which is important in facilitating good decisions (e.g., potentially affected areas, severity of disaster, evacuation orders) as the crisis unfolds. In order to help these decision makers navigate through these data, there is a need for an advanced information-sharing and interaction system that is reliable during a disaster and versatile for heterogeneous data sets. The goal of this project is to develop a multi-site visualization tool that can support disaster decision making by integrating and combining data from different sources. There is a high dependence on a reliable, flexible IT infrastructure in order to maintain the continuity of data viewed by the decision makers and the interactivity between decision makers, in order to facilitate information exchange and proper decision making. The IT infrastructure must respond and adapt to demands from users and applications as well as network resources, which dramatically changes when a disaster strikes.

**Disaster Information Management**

Research and development of a big data, information interface for disaster management professionals

Natural disasters are global events that affect multiple communities and have no international boundaries. Because of this, our disaster information management project requires collaboration on a global scale to address the problem of providing timely, accurate, and continuous data to authorities and decision-makers during a natural disaster. Open data and technologies are incorporated in this information interface to provide a solution for disaster management professionals.

**Multi-site Visualization Interface**

To address these needs, we are researching and developing a virtual “whiteboard” or workspace where the same data is presented to multiple visualization environments in geographically distinct regions. We are leveraging on SAGE2 (Scalable Amplified Group Environment) from the University of Illinois, Chicago and the University of Hawaii, Manoa; which provides an environment that allows people to simultaneously interact with visualized data located in geographically-dispersed locations on a single large-scale, high-resolution display.

The visualization interface utilizes a data broker to parse data from different sources and convert it into coordinate-specific JSON data that can be interactively displayed on a geographical map. Sources of data include PostgreSQL databases, URLs, local files, and even directly from a sensor itself.

Currently, we are working to integrate the experimental SDN PRAGMA-ENT testbed with this interface to explore concepts of reliability and resilience of the infrastructure to maintain continuous information on the visualization interface. One of the persistent challenges is identifying new open data sources to test the robustness of the data broker and to further examine issues of data integration and analysis, especially during rapidly changing events of a disaster. Finally, user interaction studies will be carried out to refine the interface and determine how usable the presented data are for disaster management professionals.