Social Value Creation Accelerated by MCSoC

Yuichi Nakamura, PhD
General Manager of Green Platform Research Labs.
NEC Corp.
NEC brings together and integrates technology and expertise to create the ICT-enabled society of tomorrow. We collaborate closely with partners and customers around the world, orchestrating each project to ensure all its parts are fine-tuned to local needs. Every day, our innovative solutions for society contribute to greater safety, security, efficiency and equality, and enable people to live brighter lives.
Agenda

1. Introduction of NEC and Social Value Creation
2. Role of Computer: from ICT to Social Value Creation
3. Hetero Computing to Contribute Social Value Creation
4. Social Value Creation Examples by Hetero MCSoC
5. Vector Processor
6. Future with Hetero Computing accelerated by MCSoC
Introduction of NEC and Social Value Creation
Is NEC a Display and Projector Maker?

NEC launches large venue laser projector: world’s most compact 5,000 lumen laser projector (Aug. 2015)

NEC Display Solutions unveils ultra-short-throw interactive projector to enhance learning environments (Jul. 2015)

NEC Display Solutions presents latest cinema innovations at CineEurope 2015 (Jun. 2015)

NEC strengthens 4K UHD professional display series with 98-inch model (2015)

No, we are ICT (Information and Communication) Company.
History of NEC's Technology

**Satellite Communication**
World 1st TV broadcast among US and Japan provided J.F. Kennedy assassination (1963)
TV broadcasts of 18th Olympiad in Tokyo (1964)

**Semiconductor**
World top share (1985~1991)
Terminated (2010)

**Carbon-nanotube**
NEC discovered a unique graphite crystal (1991)

**Super Computer**
“The Earth Simulator”, the world’s fastest super computer (2002~2004)
SX-ACE (2014)

**Hayabusa (satellite)**
Succeeded in bringing asteroid samples back to Earth (2010)

**Biometrics**
NEC's NeoFace® Facial Recognition Technology Ranks First in NIST Testing for Third Consecutive Time (2014)
The Earth in 2050

Energy demand: 1.8 times
Greenhouse gas: 1.5 times
Demand for food: 1.7 times
Demand for water: 1.6 times

Urban population: 6.3 billion people (currently 3.5 billion people)

© NEC Corporation 2015
We have to solve various problems in the world.

Solve the Social Problems

ICT is only the key to solve the problems.
Value = Change the social and human life

Safety 安全  Security 安心  Efficiency 効率  Equality 公平

Sustainable Earth  Safer Cities & Public Services  Lifeline Infrastructure  Communication  Industry Eco-System  Work Style  Quality of Life

Orchestrating a brighter world
Role of Computer: from ICT to Social Value Creation

1. Introduction of NEC and Social Value Creation
2. **Role of Computer: from ICT to Social Value Creation**
3. Hetero Computing to Contribute Social Value Creation
4. Social Value Creation Examples by Accelerations
5. NEC's Vector Processor
6. Future with Hetero Computing
For innovation of social infrastructure, high performance computer platform is required to satisfy total social optimality and individual hospitality.

**Total social optimality**

- Power
- Water
- Traffic

**Novel information communication networks**

- Optimized operations
- Optimized allocation of CPU/NW resources

**Individual hospitality**

- Aggregation and utilization of various sensor information scattered on infrastructures

Power transmission network

Water transmission network

Railway network

Road network
Fast Operation is Very Good!

Fast operation can change the world. From “visualization” to “prevention”

Severance video application

Analyzing in 1 hour
Specify Criminal
We could capture him.
Visualization

Key Device of the various kinds, fast process
Hetero MCSoC

Analyzing in 1 min
Specify the sign of crime
We can prevent crime
Prevention(Control)
Hetero Computing
to Contribute Social Value Creation

1. Introduction of NEC
2. Role of Computer: from ICT to Social Value Creation
3. **Hetero Computing to Contribute Social Value Creation**
4. Social Value Creation Examples by Hetero MCSoC
5. NEC's Vector Processor
6. Future with Hetero Computing
Computing Trend

Limit of Moor's law:
more transistors != higher performance

Single-core (increasing Hz)

Power wall

Multicore/Manycore
(increasing # of cores)

Dark Silicon

Hetero Computing
(use of accelerators)

http://www.gotw.ca/publications/concurrency-ddj.htm

http://newsroom.intel.com/docs/DOC-3126
Hetero Computing

Combined use of multiple types of engines for higher performance beyond limitation (performance and power consumption) of homogeneous computing

Engines for hetero computing

- von Neumann Architecture
  - CPU
  - GPU
  - Manycore
  - Vector Processor
- non von Neumann Architecture
  - FPGA
Select the best processing engine (accelerator) for the target application.
- The most important thing is to know the advantages and the disadvantages of the processing engines.
- The length of thread, the length of data size, the kind of processing, and etc.
- How to implement the application to the processing engine

Combine the selected processing engines (not presented today)
- Communication architecture
- Scheduling
CPU/GPGPU/Manycore/Vector: Serial program is divided to fit the processing units.

Serial Program

Parallel Processing

Parallel Processing

loop iterations

iterations are processed in parallel

Some Complex

Many Simple

Size of a core = Program complexity
SW view of non von Neumann architecture

FPGA : flexible processing unit → difficult to use

- MS, JP-Morgan, and Baidu use FPGA for bigdata processing and machine learning

Serial Program

Parallel Processing

Parallel Processing
Example (1/3): Manycore

**Manycore Accelerator**
- Many number of simple cores (ex: 80 cores)

**Example: Intel Xeon Phi series**

- Programmer can use an accelerator as a standalone node with multiple cores
  - OS (Linux) is running on an accelerator. Programmer can run a whole program on it.
- Programmer can also use it to run specific parts of a program by using dedicated compiler

```
#pragma offload target(mic)
#pragma omp parallel for for( i=0; i< N; i++) {
    ...
}
```

- Use of accelerator
- Loop parallelization using OpenMP API
- Data transfer can be written in a pragma
Example (2/3): GPGPU

GPGPU (General-purpose computing on graphics processing units)
- GPU: Processor for graphics. Widely used in standard PC
- First application: real-time 3D graphics
- Difficult: communication among CPU, GPGPU and Memory

Example: NVIDIA CUDA

From CUDA C Programming Guide Version 7.5
Example (3/3): FPGA

**FPGA (Field-Programmable Gate Array)**

- HW that users can program its HW logic
- Network card with FPGA (FPGA-NIC), CPU with FPGA (Intel announced). Microsoft, Baidu, and etc. use it for big data applications.
- Very high speed with fine grain logic level parallelism. But very difficult programming with special knowledge of HW logic design.

Programming tool (HLS) for HW design is very important (Ex: NEC's CyberWorkBench supports C programming)

---

**HTG Xilinx Virtex6 PCIe Gen2/SFP+/USB 3.0 DevBoard**

http://hitechglobal.com/Boards/Virtex6_PCIExpress_Board.htm

writable logic element (CLB)
Performance Improvement by Selecting Optimal Processing Engines

The 100 times better performance than general-purpose processors can be achieved by selecting optimal processing engines and leveraging their performance.

By limiting the use to specific processes, the performance far exceeds general-purpose processors.

- **Separation processing of huge data**
- **Collective processing of huge data**

Vector processing engines:
- 100x performance by using many-core
- 100x performance by using vector

Many-core processing engines:
- 100x performance by using many-core

General-purpose processors:
- By limiting the use to specific processes, the performance far exceeds general-purpose processors.

Biometrics
Earthquake simulation
Combination of Processing Engines

- Social system is consisted of various applications.
- Each processing engine has an advantage and disadvantage according to application/task/process/function.
- One engine cannot solve the problem.
- Use the best processing engine and combine them.

Social System Control

Collecting ▶ Analysis ▶ Control

Xeon Server

Many Core

GPGPU/Phi

FPGA

Tasks allocate the best accelerators/Engine

Real time Constraint

Example: Video severance

Each engine has an advantage and an disadvantage

100X faster than a single server
The performance is maximized by optimizing the combination of heterogeneous processing engines with software technologies to achieve the required processing capability at low cost.
ExpEther: How to connect many processing engines

ExpEther provide a single-hop PCI Express switch virtually over Ethernet.
- combination of up/downstream EE bridge and Ethernet transport.
- Ethernet is transparent to the CPU/OS/software.
- Equivalent to a single-hop standard PCIe switch.

Utilize commercially available server, OS, I/O, driver w/o modification.
Social Value Creation Examples by Hetero MCSoC

1. Introduction of NEC
2. Role of Computer: from ICT to Social Value Creation
3. Hetero Computing to Contribute Social Value Creation
4. **Social Value Creation Examples by Hetero MCSoC**
5. NEC's Vector Processor
6. Future with Hetero Computing
Our experience: History of MCSoC from NEC

Application Specific
- Image Recognition
- NW Contents Processing
- Programmable Hardware
- Motion Estimation of Objects around Vehicles
- Simultaneous Interpretation (Restricted)
- Speech Recognition (100K words)
- AV CODEC (Restricted)

General Purpose
- Merlot
- MP211
- MPCore
- Medity-1/2

2010

make

use

GOPS / GIPS

Manycore

GPU

FPGA

Application Specific Architecture

X 40 Times

General Embedded

Year

http://newsroom.intel.com/docs/DOC-3126
http://hitechglobal.com/Boards/Virtex6_PCIExpress_Board.htm
http://techon.nikkeibp.co.jp/english/NEWS_EN/20090918/175471/?SS=imgview&FD=1825605759&ad_q&rt=nocnt
Merlot (prototype multiprocessor for embedded system)

World 1st multiprocessor with high performance and low power (ISSCC2000 Highlights)

- **On-chip parallel architecture**
  2way super scaler, 4PE, 128KB

- **Standard C programming**
  Automatic parallelization of C

High performance and low power

- 1GIPS 1W prototype, Merlot[0.15μm]

NEC developed multiprocessor products, MP211 and MPCore etc, based on Merlot

I joined this project as only the design tool member.
MPCore (SMP for embedded system)

MPCore TC (2005)
- ARM11 Core x4
- coherent distribute L1 cache
- distributed interrupt controller
- shared L2 cache (1MB)

NEC provided the technologies of MP to ARM.
Technologies for MCSoC/Hetero Computing based Accelerations

- Technologies to make a parallel processor
  - Power management technologies such as DVFS
  - HW and SW support for parallel processing such as specialized cache for control-flow parallel processing

- Technologies to use various parallel processor or accelerators
  - Parallelization of SW for efficient use of MCSoC
  - Co-working of different type of processor in a hetero system
  - Algorithm design for various accelerators.

Software for Serial -> Hetero MCSoC -> Customized Software For Accelerators
Performance

Load Balancing

Parallelization

Manycore System

Program

SW Technology (Parallelization & vectorization)

Extracting parallelism for multiple cores

Scheduling to keep all cores busy
Example 1: Face Detection for Large-scale Video Surveillance

- Real-time face detection for large-scale surveillance system such as Stadium
  - Detecting hundreds of faces in a 4500x3000 image in 1 second

Process:
- Input
- Candidate extraction
- Candidate evaluation
- Output

Serial version (standard server):
- Serial execution on 1 core
- ~100 sec

Parallel version (Manycore system):
- Parallel execution on hundreds of cores
- ~1 sec
Extraction of Parallelism

Parallelization 1: Parallelization in each stage
● parallelism among divided images (stage 1), candidates (stage 2)

Parallelization 2: Parallelization among stages
● improving overall parallelism to use many cores with low overhead

candidate extraction
parallelism among regions
candidate evaluation
parallelism among candidates
output
Load Balancing

Problem: Number of faces in a region varies depending on situation
  • static distribution of regions to cores causes load unbalance

Solution: developing dynamic load balancing
Standard Server with Developed SW Technologies
Example 2: GPGPU Realtime Video Codec

Offloading ALL image processes to GPU for reduction of CPU-GPU Communication

Parallelization technologies for using thousands cores

- Divide into hundreds parallel tasks
- Spatial & temporal parallelization
Super Resolution is an image processing to improve image quality (not realtime on a standard server).

Realtime SR on a GPGPU server enables detection of suspicious persons in unclear camera image.

**before super resolution**
(noisy image from highly sensitive camera)

**after super resolution**
Calculating NK225 index in realtime for streams from TSE

Helps to make decisions earlier than software-users

- Realtime stream of stock prices
- FPGA is used for only index calculation
- Speed up by 3.3x
FPGA Compiler Technology

- FPGA based method can reduce the processing time
  - But, hardware design time is larger than software design time.
- Stock trading algorithm is changed everyday.
  - Fast design tools for FPGA based method is required.

Stock Trading Algorithm written in C/C++/event lang (SQL).

MS and JP Morgan: Implementing by hands

30min for One change

our featured point

- financial
- networking
- smart grids
- factories/plants
- FPGA NIC
- event processing

Compiler
NEC's Vector Processor

1. Introduction of NEC
2. Role of Computer: from ICT to Social Value Creation
3. Hetero Computing to Contribute Social Value Creation
4. Social Value Creation Examples by Hetero MCSoC
5. **NEC's Vector Processor**
6. Future with Hetero Computing
NEC has always provided the high sustained performance by Vector Super-Computer SX series.

**World No.1 Super Computer (Top500 2002~2004)**

NEC has always provided the high sustained performance by Vector Super-Computer SX series.

**Gordon Bell Award (2002)**

Support >1000 nodes
ECO
ES2

**Support Multilane IXS**

1990 2000 2010

**SX-2**

Water Cooling
Vector Processor

Scalar

SIMD (Modern Scalar)

Vector(SX)
256 elements are processed per instruction

Vector is efficient to fill pipelines and to hide latencies
According to Japanese Government (MEXT) working group report for a wide variety of strategic segment applications, diverse characteristics are observed. MEXT: Ministry of Education, Culture, Sports, Science & Technology

B/F requirement from each application differs greatly. Any single architecture cannot cover all application areas.

Processor Overview

**CORE**
- Architecture: Vector
- Clock Frequency: 1.0GHz
- SPU decode rate: 4 instructions
- VPU Performance: 64GFlops
- ADB size: 1MB
- ADB bandwidth: 256GB/s
- Memory bandwidth: 64GB/s~256GB/s
- Core Byte/Flop: 1.0 ~ 4.0

**CPU**
- Cores: 4
- Performance: 256GFlops
- Memory bandwidth: 256GB/s
- CPU Byte/Flop: 1.0
- Memory capacity: 64GB
Single Core Comparison

The SX-ACE core can provide the world top-level performance and the largest memory bandwidth.
Vector processor has a great advantage to process streaming huge data continuously.

Bigdata which collects many terminals, consists of “streaming huge data”.
- All data unit are applied the similar processing.

We confirmed our vector processor can achieve significant speed up compared with CPUs or Many core processors, on several bigdata processing.

Bigdata Processing with Vector Processor

Water supply management

- Memory intensive
- Vector Processor
- Streaming huge data at once

More than 100x performance by using vector
Aurora for Big Data

Doing same computation for huge amount of data is home field of vector processor

Current target

- Weather forecast
- Climate change
- Air craft, Automobile
- New energy
- New material
- Health-care
- Societal analysis

Expanded target

- Big data
- Image analysis
- Financial simulation
- Machinery Architecture
- Resource exploration
- Biotechnology (Drug discovery)
- Disaster prevention/mitigation
- Defense (cryptanalysis)
- Financial simulation
- New energy
- New material
- Health-care
- Societal analysis

Big data

-气象予報
- 気候変動
- 飛行機、自動車
- 新エネルギー
- 新素材
- 医療
- 社会分析
- 防災/対策
- 新エネルギー
- 財政シミュレーション
- 防守
(暗号解析)
Future with Hetero Computing

1. Introduction of NEC
2. Role of Computer: from ICT to Social Value Creation
3. Hetero Computing to Contribute Social Value Creation
4. Social Value Creation Examples by Hetero MCSoC
5. NEC's Vector Processor
6. **Future with Hetero Computing**
Problem Toward Efficient Real World System

Real world = Mixture of distributed many types of devices

Affinity of Real World and Distributed Hetero

- Sensor
- Edge
- Actuators

Smart Water System
Optimal Control of Real World System

Allocation and scheduling of large-scale system (~100K)

Optimal control (scheduling and allocation = Nearly equal technologies of MCSoc) of large-scale real world system by ICT can generate new value.
ICT systems can change the world to realize social creation value creation.

Hetero MCSoc is a key device of social value creation.

There are various types of application in the real world. Then, we have to select the best acceleration or MCSoc according to the applications.

The development of MCSoc is very important. However, the use of MCSoc efficiently is also important.

In the next stage, integrated systems by using many MCSocs are one of main streams of ICT systems to contribute social value creation.

A basic technology of integrated MCSocs is the same as the design technology of MCSoc, allocation and scheduling.

Let’s contribute to social systems by MCSoc technology!
Orchestrating a brighter world

NEC