OCTOPUS: a new supercomputing service of Osaka University

Susumu Date Cybermedia Center, Osaka University

Cybermedia Center, Osaka University

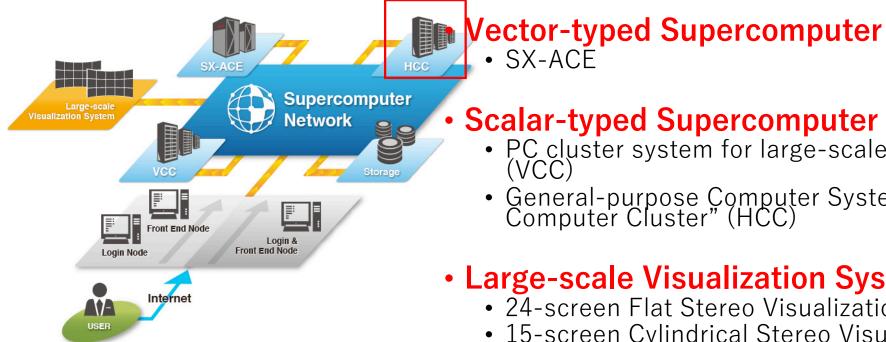
Cybermedia Center

- a computing center at Osaka University
 - <u>http://www.cmc.osaka-u.ac.jp/</u>
- has a responsibility of providing a powerful high-performance computing environment for university researchers across Japan as a national joint-use facilities.
- Is in charge of the campus IT infrastructure and for promoting its effective use.



IT Core Annex: new datacenter building

Overview and Feature of Cybermedia Center's Computing Infrastructure



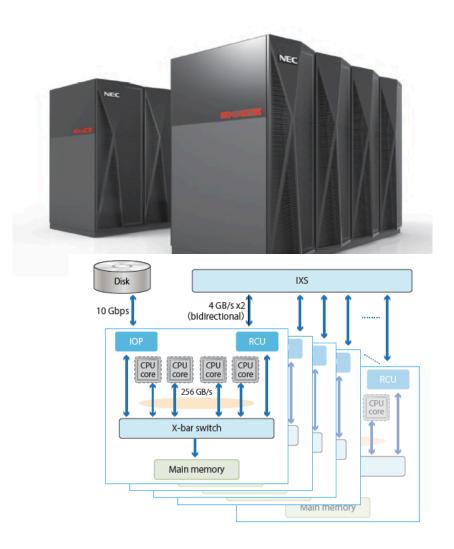
Scalar-typed Supercomputer

- PC cluster system for large-scale visualization (VCC)
- General-purpose Computer System "Hanyou" Computer Cluster" (HCC)

Large-scale Visualization System

- 24-screen Flat Stereo Visualization System
- 15-screen Cylindrical Stereo Visualization System

SX-ACE of Osaka University



Type: Vector OS: Super-UX # of nodes: 1,536 # of cores: 6,144 Total memory: 96 TB Peak performance: 423 Tflops (inc. scalar processor)

Each node has

- 4-core multi-core vector CPU (256 GFlops)
- 64 GB memory

IXS (Internode crossbar switch) connects512 nodes and allows 4GB x 2 communication

VCC(PC cluster system for large-scale visualization)



Type: Scalar OS: Linux # of nodes: 66+3 # of cores: 1320 + 84 Total memory: 4.224 + 0.192 TB Peak performance: 26.4 + 4.5 TFlops Accelerator: NIVIDA Tesla K20 x 59

• As of 2017.11

66 comptuing nodes (Intel Xeon E5-2670v2 processor, 64GB main memory)

+

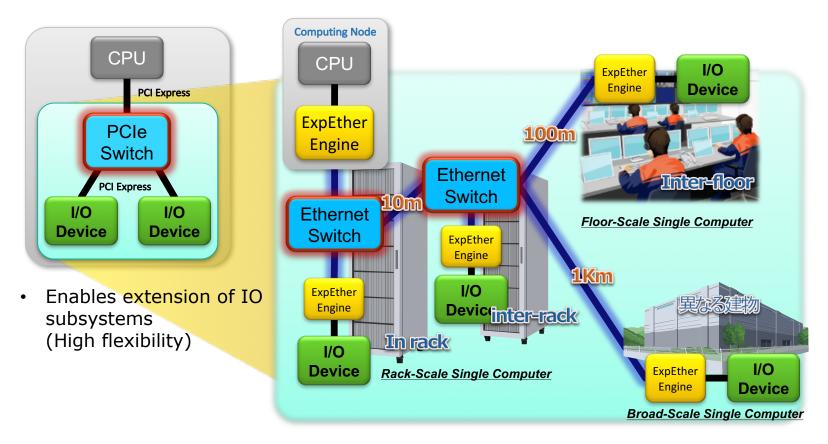
3 computing nodes (Xeon E5-2690v4, 64GB)

• Dynamically Reconfigurable Cluster System using ExpEther, system virtualization technology from NEC

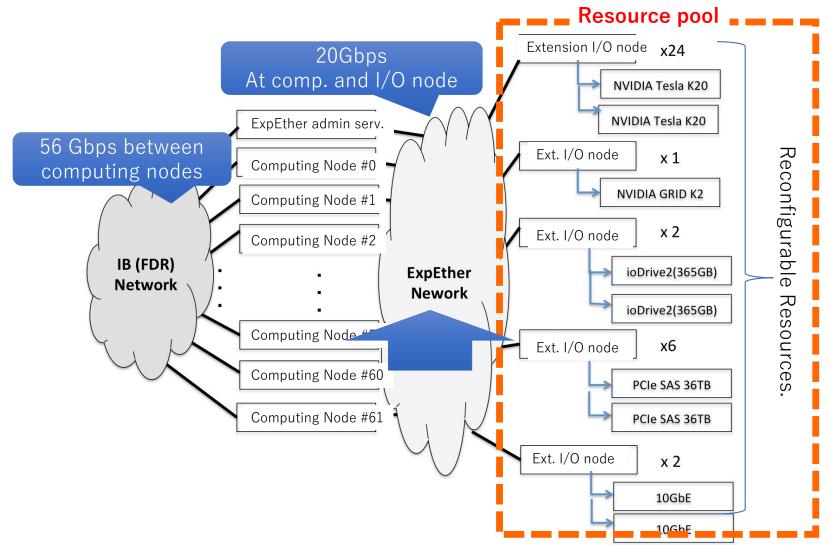
NEC slide

ExpEther ?

- The world's first technology that extends PCI Express onto Ethernet (L2 layer)
 - Enables "spacious single computer" using networks"



Key feature: Reconfigurability



General-Purpose Computing Cluster "Hanyou" computing Cluster" (HCC)



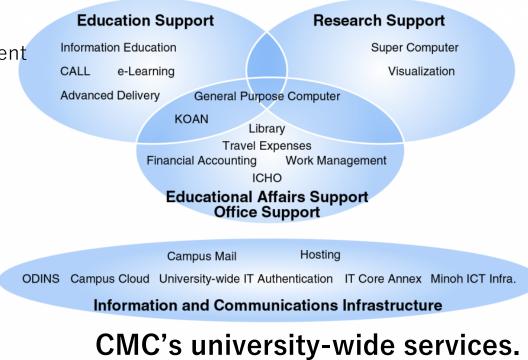
Type: Scalar (VM) OS: Linux # of nodes: 575 # of cores: 1150 Total memory: 2.6TB Peak performance: 16.6 TFlops Expectation to a new scalar system (Background behind system procurement)

Problem of our scalar-typed computing resources keeping users away from the CMC

- Our scalar-typed computing resources are not sufficient enough to satisfy users requirements and needs.
- The problems of each system is caused from:
 - HCC was procured to be not a high-performance computing resource but "general-purpose" for university-wide .
 - VCC was procured under the supplementary budget from the government, which results in higher service charge.
 - The supplementary budget is expected to use up within the fiscal year and so the future maintenance and support fee cannot be covered.

Problem of HCC system

- In the last procurement, HCC was procured as a part of "general-purpose computer" composed of Educational Computer system, CALL system, High-Performance Computer system, Library system.
 - Other education-related systems compresses HPC system.
 - The budget was integrated for the procurement
 - Education, Library >> HPC
 - University is Education!!
 - Online Journals are more important..



Problem of HCC system

- In the last procurement, HCC was procured as a part of "general-purpose computer" composed of Educational Computer system, CALL system, High-Performance Computer system, Library system.
 - Other education-related systems compresses HPC system.
 - The budget was integrated for the procurement
 - Education, Library >> HPC
 - University is Education!!
 - Online Journals are more important..



Windows+ parallels HW (CPU) IB + network HPC system? Why does MPI communication error take place?

when I moved mouse.

- HCC is no longer a system for high-performance computing but education system.
 - Unstable due to virtualization. (Parallels stop support.. Unknown behaviour)
 - Not high-performance.

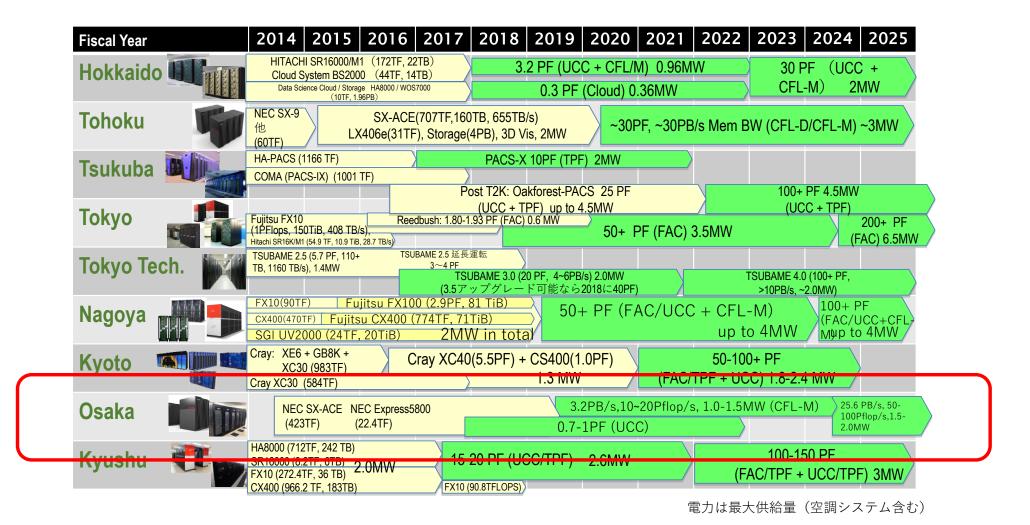


Users "hope" has become "hopeless" ...

Towards OCTOPUS procurement

- Happy "divorce" from education system, CALL system and Library system
 - "property division" based on original budget for each system (no compentation, no child support : becomes others!)
- re-review users' requests to next supercomputer & recover users for scalar-typed systems.
 - Towards hybrid type of cluster system that can accommodate a diversity of user requirements.
 - No vector for this procurement. Procurement timing is off. New vector arch. Is too late for our procurement.
 - Dig the potential users' needs to large-scale memory machine
 - Rescue users who used to enjoy SX-9 with large-memory
 - Life scientists are eager to use.
- Towards large-scale system that satisfies computing needs to scalar-typed computing.
 - Number of nodes > number of cores, number of GPUs
 - Storage 3PB: ScateFS did not support 100Gbps interconnect.
 - Over 1 Peta Flops
 - CMC's total peak performance of all systems is less than 0.5 PFlops although HPC people say post-peta, exa.. (efficienty is important though)
- Stable performance supply!
 - Water cooling due to higher CPU, GPU TDP
 - Our datacenter building has enough water cooling facility.
- Open "procurement", of course !
 - Is NEC only looking to JAMSTEC and Tohoku University?

System planning of 9 computing centers (As of May 2016)



Real Voice from Users(1)

I hope next system have more computing nodes because waiting time is long due to higher utilization.



A system that allows us to perform middle-class parallel computing (nonetheless, parallelism should be around thousands) towards the future use of K computer.

Vector typed system is preferable.



As I am planning to use my own code (simulation), I don't have any requests to application installation. My first choice is a cluster system where MPI and OpenMP are available. In near future, I would like to consider the use of GPU and FPGA. I do not recommend any vector-typed machine due to lack of generality.

I prefer x86 server with large amount of memory. Pascal-GPGPU is also preferable。



I am strongly hoping that the CMC should have more application software installed. In comparison with other computing centers in the former imperial universities, application software is too weak.



Real Voice from Users(2)



It is preferable that memory capacity is more than 64GB (the same size of SX-ACE), 128GB if possible. Size of recent \blacksquare is becoming large and so I hope next system has enough memory. Memory access performance is also important.



Please put higher priority to communication performance. (I hear the story of $\bigcirc X$ university. Because they squeezed a penny for communication performance, efficiency of \blacksquare \blacksquare code had dropped to 0.1 %.) I hear other similar stories of taking communication lighter to raise peak performance. Again, do not squeeze a penny for communication.



I personally prefer NVIDIA GPU because it offers higher perormance. (*snip*) In the case of introducing GPU to the next system, please provide user training sessions about CUDA and OpenACC.



XeonPhi. I hear bad stories about Knights Corner, the first generation available now (OOuniversity) I would like to expect next-generation Knights Landing.. For I user, XeonPhi may be better than NVIDIA GPU because users can use it without a lot of knowledge and expertise.

Real Voice from Users(3)



I hope that next system should have more computing nodes, because waiting time is too long and it seems that many users are waiting.

Al "tensorflow", a python module must have glibc v.2.15 later and so, for example, newer version CentOS 7 later would come in handy



I am doing the research of utulizing plasma simulator at $\bigcirc X$ science research institute. I would like to use a cluster system with more tens of thousands cores.

My specialty area is fluid engineering. Recently, it is natural to perform parallel computing using multiple cores. I feel like "consuming computing resources as infrastructure" the same way as water and electricity, rather than "using computing resources in a trial-and-error manner so far. Although I misunderstand because I am directly not using the computer systems at the CMC, I often hear (a) MPI job is not executed, (b) system often stops, (c) data staging is always required all the time. These situations are bad and inconvenient as computing infrastructure. I would like to request the improvement of these situations.



Roadmap to OCTOPUS

仕様書

全国共同利用大規模並列計算システム

Large-scale Parallel Computing System for Nation-wide Shared Use

大阪大学サイバーメディアセンター 平成**29**年**2**月

- Feb.17, 2016 Explanatory meeting of system introduction
- Jul.28, 2016 Explanatory meeting of specification draft
- Feb. 6,2017 Bid announcement
- Feb. 13, 2017 Pre-bid meeting
- Apr. 3, 2017 Deadline for bid
- May 31, 2017 Bid Opening
- Dec. 1, 2017 Service starts

OCTOPUS Overview

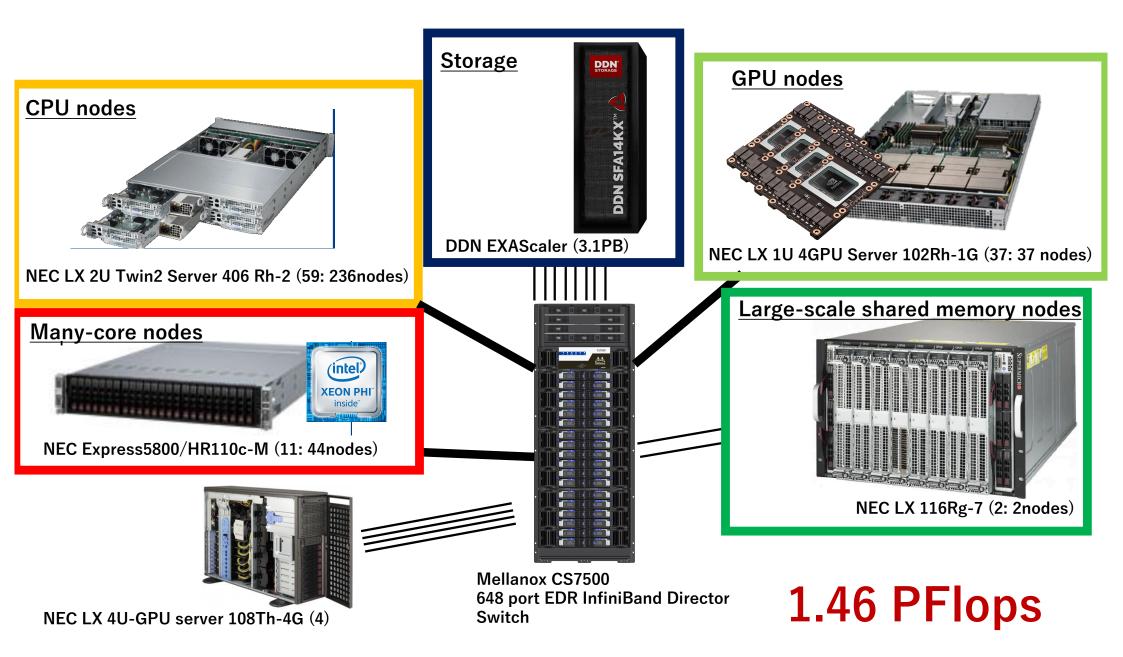
OCTOPUS

- is short for Osaka university Cybermedia cenTer's Over-Petascale Universal Supercomputer.
- We are really eager to obtain a Peta-Flops system.
- The CMC' system is difficult to remember. Not fun!
 - Vcc, ace, hcc.. PCC.
- Many Japanese people associate TAKOYAKI with Osaka.

Takoyaki (たこ焼き or 蛸焼) is a ball-shaped Japanese snack made of a <u>wheat flour</u>-based <u>batter</u> and cooked in a special molded pan. It is typically filled with minced or diced <u>octopus</u> (*tako*), <u>tempura</u> scraps (*tenkasu*), <u>pickled ginger</u>, and <u>green onion</u>.^{[1][2]} Takoyaki are brushed with takoyaki sauce (similar to <u>Worcestershire sauce</u>) and <u>mayonnaise</u>, and then sprinkled with green laver (*aonori*) and <u>shavings</u> of dried <u>bonito</u>. There are many variations to the takoyaki recipe, for example, <u>ponzu</u> (soy sauce with <u>dashi</u> and citrus vinegar), goma-dare (sesame-and-vinegar sauce) or vinegared dashi.

Yaki is derived from "yaku" (焼く) which is one of the cooking methods in Japanese cuisine, meaning "to fry or grill", and can be found in the names of other <u>Japanese cuisine</u> items such as <u>okonomiyaki</u> and <u>ikayaki</u> (other famous Osakan dishes) from wikipedia





NEW ARRIVAL CYBERMEDIA CENTER. DSAKA UNIVERSITY SUPERCOMPUTER "OCTOPUS" 全国共同利用大規模並列計算システム、12月新規導入。



大阪大学 サイバーメディアセンター 応用情報システム研究部門 大阪大学 情報改進部情報基盤課 研究系システムモ systemのcmc.osake-u.ec.in. 05-6879-8813 2017年 12月初旬 サイバーメディアセンター 利用者向け 先行教育講談

Procurement results

NEC fielding awarded.

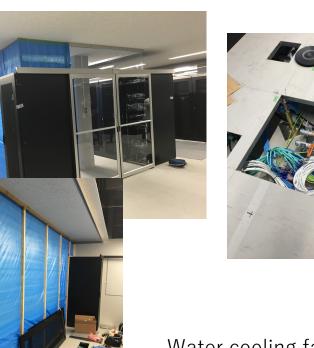
- A hybrid cluster system composed of
 - General CPU nodes: 236 (471.24TFlops)
 - Intel Xeon Gold 6126 (Skylake/ 2.6Ghz, 12 cores) x 2
 - 192GB memory
 - GPU nodes: 37 (858.28TFlops)
 - Intel Xeon Gold 6126 (Skylake/ 2.6Ghz, 12 cores) x 2
 - 192GB memory
 - NVIDIA Tesla P100 (NVLINK) x 4
 - Many-core nodes: 44 (117.14TFlops)
 - Intel Xeon Phi 7210 (KNL/ 1.3Ghz, 64 cores) x 1
 - 192GB memory
 - Large-scale shared-memory nodes : 2 (16.38 Tflops)
 - Intel Xeon Platinum 8153 (Skylake/ 2.0Ghz, 16 cores) x
 8
 - 6TB memory
 - Storage (Lustre 3.1 PB)

However, current situation is....

- **Delayed!!!!** Waiting for the arrival of CPU/GPU/many-core/largememory nodes....
 - Right before leaving Japan, I hear the first server is scheduled to arrive on Oct. 28.



DDN Storage and Mellanox was installed on Sep.11.





Water cooling facility has been already expanded in our data center building.

Cybermedia Center Osaka University

Basic policy/strategy for supercomputing system planning

- Targeting high-memory bandwidth and so aims to improve 8-10 times larger in memory BW every procurement
- Promotion of cluster architecture that can allows users to use vector and scalar computing depending on applications.
- Streamlining the entire scientific workflow from data collection to visualization

Fis	scal Year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	L
Current		(men	SX-ACE hory BV C (type:	V393.2	TB/s, p	eak per			s, powe	r consi	umptior	n 0.74N	/\\\)	
next					UCC	nemory		2PB/s? – 1PFlc		perf. 1	0-20P1	lops, p	ower 1.	0-1.5MW)
Next next		and a	olloba	ration										00Pflop,

Expansion of industrial use and collaboration

- From social contribution and regional contribution, we would like to actively conduct collaboration and industrial use
- Seeking for New computing needs

Human resource development

- Nurturing of human resources who can solve e-scientific problems leveraging HPC and VIS
- Development of human resources who can lead to new architecture based on integrated vector and scholar

Towards a new computing service through OCTOPUS

R&D of new technology that accommodate a diversity of computing needs w/ NEC collaboration research division of CMC

- Secure Stage-in mechanism that allows scientists to treat security-sensitive data on supercomputers (last year: wssp by me)
- Flexible supercomputing that allows reconfiguration of hardware and software based on users's computing requirements (two years ago: wssp by Prof. Shinji Shimojo)

More seminar and training opportunities to users.

- GPU challenge w/ NVIDIA
- Tuning support w/ NEC

Outreach and advertisement to new users

- Octopus rack design
- More explanatory opportunities to industry users.

Research and Development

Severity Measurement

Index of Orthodontic Treatment Priority (IOTN)

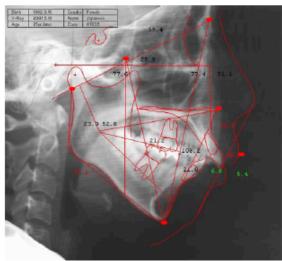
one of the severity measures for malocclusion and jaw abnormality, which determines whether orthodontic treatment is necessary

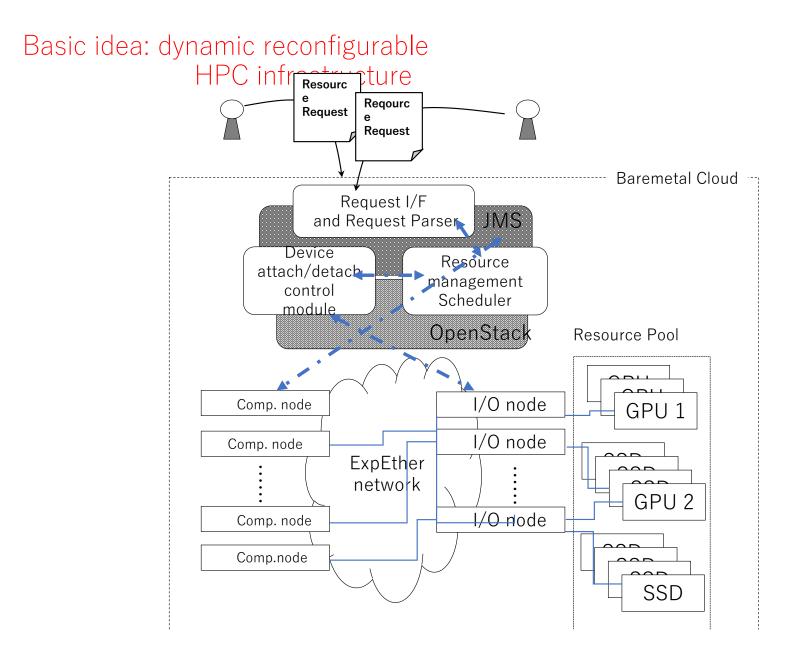


Morphological Landmarking

- Feature point detection
- measurement

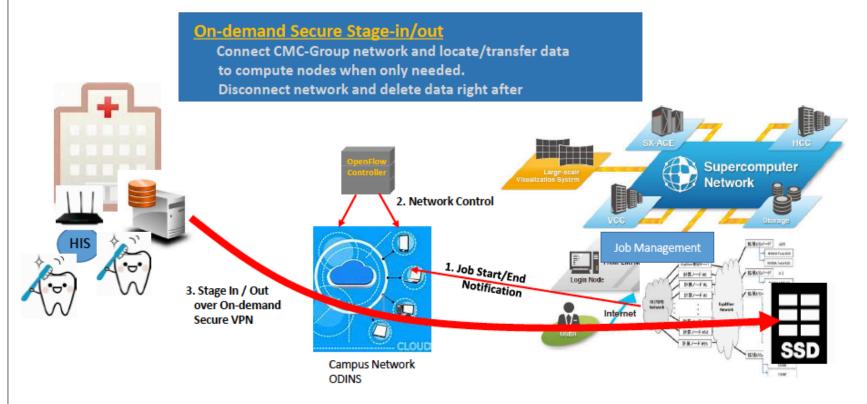
Seiya Murata, Chonho Lee, Yujin Shimizu, Chihiro Tanikawa, Kazunori Nozaki, Susumu Date, "Towards a Fully Automated Diagnostic System for Orthodontic Treatment in Dentistry", The thirteenth IEEE eScience Conference (e-science2017), Auckland, New Zealand, Oct. 2017.





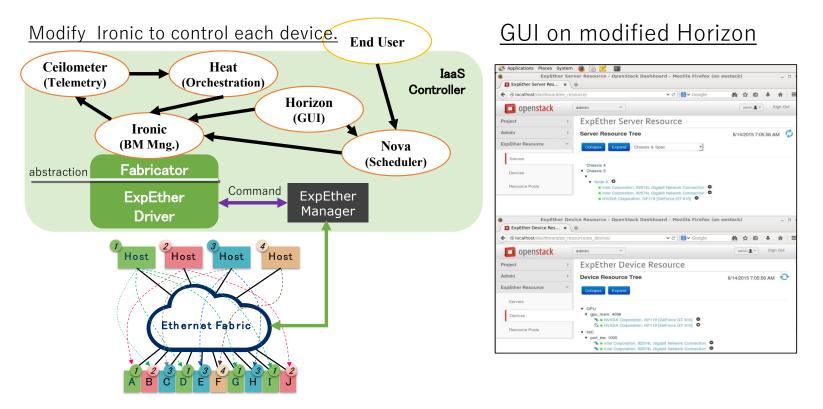
Secure VPN for Healthcare Data Analytics

 A Healthcare Infrastructure that supports secure, big data analytics using various data analytics services



Hi-laaS: OpenStack Based Resource Management

Modify Ironic (Bare metal control) to device level.



Akihito Misawa, Susumu Date, Keichi Takahashi, Takashi Yoshikawa, Masahiko Takahashi, Masaki Kan, Yasuhiro Watashiba, Yoshiyuki Kido, Chonho Lee, Shinji Shimojo, "Highly Reconfigurable Computing Platform for High Performance Computing Infrastructure as a Service: Hi-IaaS", The 7th International Conference on Cloud Computing and Services Science (CLOSER 2017), pp. 135-146, Apr. 2017. [DOI: 10.5220/0006302501630174]

More seminar and training opportunities attracting users

OCTOPUS Celebration GPU Challenge

or the Fu	or the Future user For the User		for Researc	ch Proposal Based Use	For HPCI	For JHPCN			
vstem Service Guide / Application Support			Support	Training/Seminar	Publication	Achievement			
		natory meetir	Ig	Cybermedia Center, Os	aka University > Ir	aining/Seminar > OC	OPUS Celei	oration GPU	Challenge
		ration GPU Chal		Cybermedia Center, Us	aka University > Ir	aining/Seminar > OC	OPUS Celei	bration GPU	Challenge

GPU computing nodes, large-scale main memory nodes, and large-scale storage. In this challenge, as a part of celebration of OCTOPUS introduction, we call for programs (noncommercial) that can be accelerated on the coming GPU computing nodes of OCTOPUS, so that GPU computing nodes are leveraged right after system introduction. The programs accepted in this challenge will be performance-tuned and optimized by the Cybermedia Center and NVIDIA experts.

GPU computing nodes of OCTOPUS has four NVIDIA GPU Tesla P100 equipped on NVLink high-speed interconnect network. This architecture provides a unified memory space of 64GB GPU memory(16GBx4) and CPU memory (192GB) for high parallelism.

This challenge aims to compare vector processing on SX-ACE and computing performance of Tesla P100. For the reason, this challenge targets programs currently executed on a single node of SX-ACE.

Please apply if you have any concern and interest to this challenge

Challenge

Applicant Guidelines	
Application Period	Oct. 2 – Oct. 15, 2017
Application Qualification	Users of the Cybermedia Center
Target Program	Programs currently executed on a single node of SX-ACE
Number of Target Programs	a few
How to apply	Please apply from the following form.

Date : Oct 2 - Oct 15, 2017
Organizer: Cybermedia Center, Osaka University, NVIDIA
Venue: -
Type : Tuning support
Quota:
Application deadline: Oct. 15, 2017

Motivation

- Many users are still stuck on a single SX-ACE node although they want to have more computational power.
 - Tuning support (migration to MPI) are available through NEC experts on the CMC though.
- Unified memory feature of P100 allows users to use 64GB GPU memory space (same as SX-ACE)
- Want to compare a single SX-ACE node a single GPU node using our user's actual code (for next procurement of supercomputer ?)
 - Performance, Programmability, Maintenability, Cost, etc..

• SX-ACE

- 4-core multi-core vector CPU 256GFlops (64GFlops x 4 cores)
- 64GB memory

OCTOPUS GPU node.

- Intel Xeon Gold 6126 (Skylake/ 2.6Ghz, 12 cores) x2 (1.9968TFlops)
- NVIDIA Tesla P100 x 4 (21.2 TFlops)
- 192GB memory

OCTOPUS Rack Design Contest

Publicly call for rack design of OCTOPUS for outreach and advertisement to the public.

スーパーコンピュータをキャンバスに絵を描いてみませんか?

大阪大学サイバーメディアセンターでは、2017年12月に日本全国の大学、研究所、企業の研究者が利用可能なスーパーコンピュータシステム 「OCTOPUS (Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer)」の運用を開始いたします。 スーパーコンピュータOCTOPUSの導入・運用開始を記念して、今後のOCTOPUSの"顔"となるデザインを募集いたします。 採用されたデザインは、本センターのデータセンタであるITコア棟内の設置エリアに配置される計算ラック群(1940mm x 3500mm)に描画され す。また、本センターのウェブページ、広報資料等、幅広く利用されます。

締切

2017年9月20日 作品提出・応募締め切り、必着

賞

最優秀賞(1点)ノートパソコン、表彰状

募集内容

「<mark>OCTOPUS</mark>」のデザイン

下記のキーワードから想像されるデザイン 「大阪」

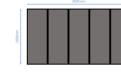
「皆に愛される」

「最先端の科学研究」

「スーパーコンピュータ」



「OCTOPUS」計算ラック群 白抜きの箇所にデザインを描画いたします



「OCTOPUS | 計算ラック群 平面図



「OCTOPUS|計算ラック群 デザイン掲載イメージ

Why don't you draw your design on the canvas on our supercomputer system?

Cybermedia Center Osaka University will start the computing service of OCTOPUS (Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer) from Dec. 2017. The supercomputer system is expected to be utilized by researchers and engineers in academia and industry. As one of celebration events, we publicly call for the rack design of OCTOPUS to be "the face of OCTOPUS system". The adopted rack designed will be put into the series of computing racks (1940mm x 3500mm) on the IT core building, our data center facility. Also, it will be widely used for the advertisement of the system through our project web page, advertisement materials and so on.

What we call for?

"OCTOPUS" rack design that can be associable with the following keywords. "Osaka"

"Beloved From Everybody"

"Cutting-edge Scientific Research"

"Supercomputer"



OCTOPUS" Computing racks. The awarded design is put onto the white region. 2D diagram



"OCTOPUS" Computing racks.

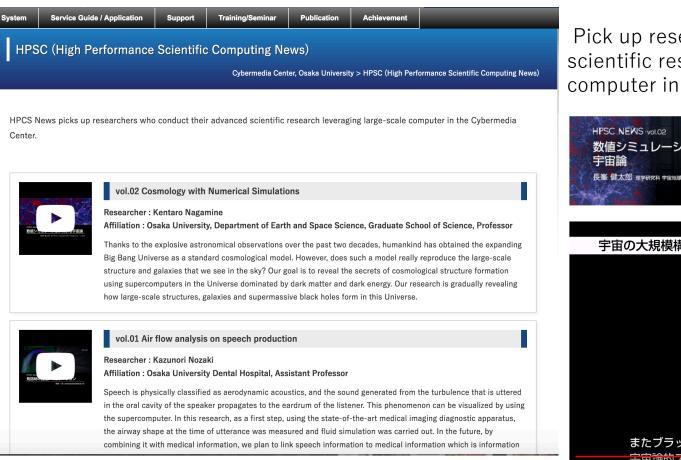


"OCTOPUS" Computing racks. Image of computing racks whether the design is put

Outreach to the public for getting the understanding of supercomputing

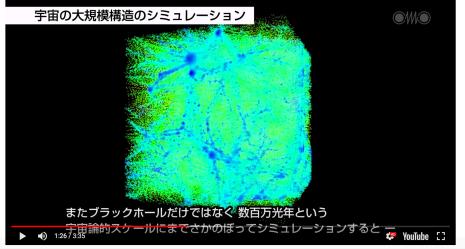
HPSC news released on Youtube.

http://www.hpc.cmc.osaka-u.ac.jp/en/hpsc-news/



Pick up researchers who conduct their advanced scientific research leveraging large-scale computer in the Cybermedia Center.





Summary

- Background (dark-side story) behind the procurement of our new supercomputer "OCTOPUS" was presented.
- OCTOPUS system is a hybrid cluster system of SKL, KNL, GPU and large-scale shared memory nodes with 3.1 PB on IB EDR. Computing service will start at Dec. 1.
- Based on how OCTOPUS is used, we are preparing for the procurement (8- times larger than OCTOPS) of next supercomputer system.