スーパコンピュータTSUBAMEとKnoppix for CUDA / OSS分野とHPC分野との接点

東京工業大学　小西史一

本日の話しの流れ

1. 発表者のこれまでの経緯について
2. Knoppixによるリマスタリングについて
3. 専用計算機から、準汎用計算機への移行
4. TSUBAME1.2について
5. OSS分野とHPC分野との接点

東京工業大学大学院情報理工研究科
グローバルCOE”計算世界観深化と展開”
特任准教授

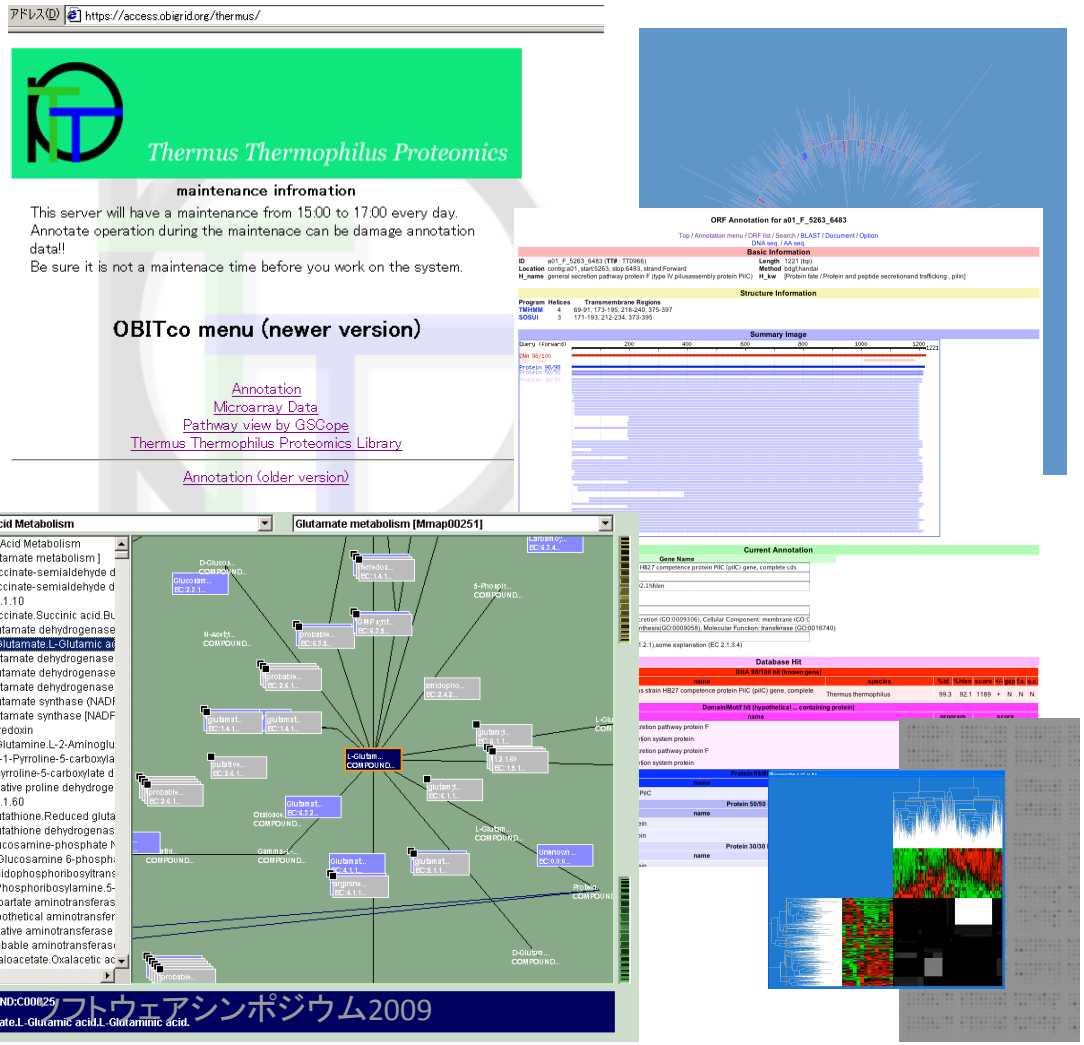
計算を中心に対象を見直すプロジェクト

独立行政法人　理化学研究所
横浜研究所　生命情報基盤部門
客員研究員

独立行政法人　理化学研究所
放射光科学総合研究センター
放射光システム生物学研究グループ
客員研究員

Thermus thermophilus whole cell simulation

Provide all of function for bioinformatics as homology search or public database service etc. And it can be simply used by a biologist.

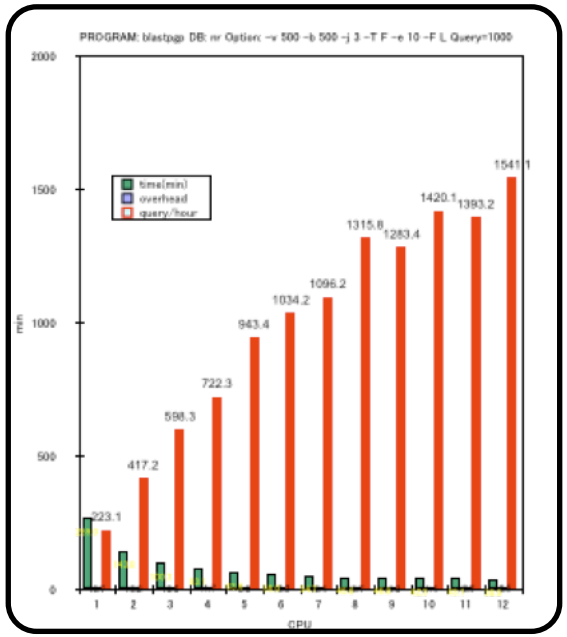


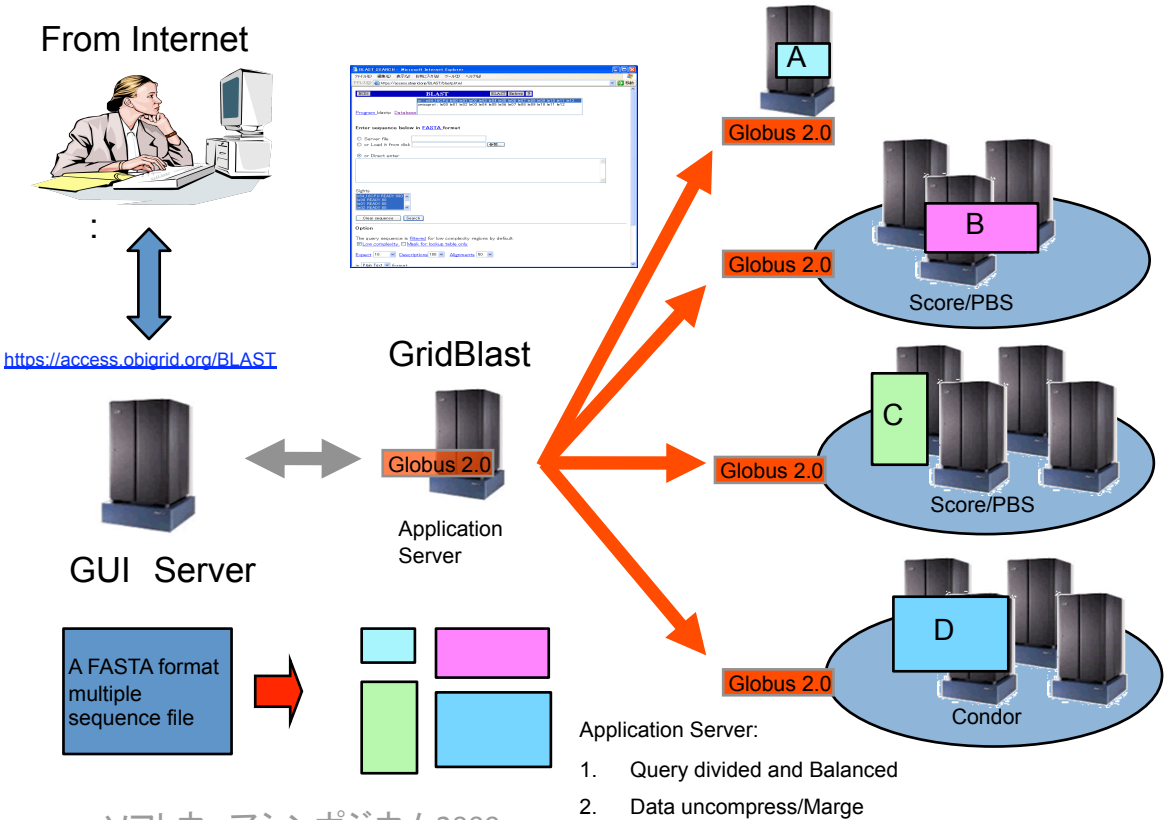
System Building for sharing knowledge about *Thermus thermophilus.*

Provide Genome information and many type of annotations that it is worth sharing as *Thermus thermophilus Research* Portal Server with secure.

GridBlast (Large Scale Homology Search)

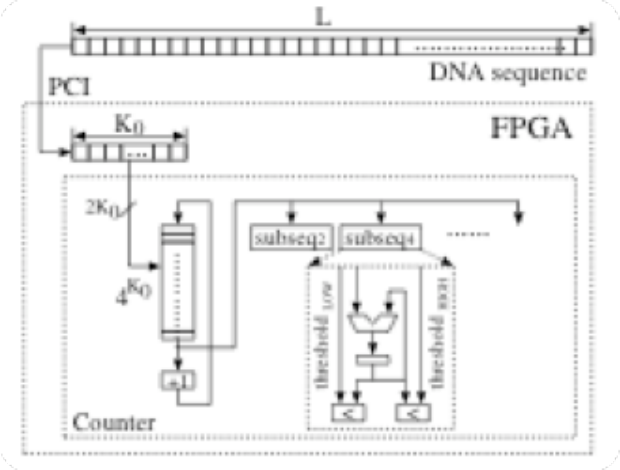
OBIGrid provides high-throughput GRIDBLAST services (OBIGbs) for researchers who need to deal with many BLAST query sequences at one time by exploiting both distributed processing and parallel processing. A new application-oriented grid framework has been introduced to split a BLAST query into independent sub-queries and to execute the sub-queries on remote personal computers and PC clusters connected by a virtual private network (VPN) over the Internet. The framework consists of five functional units: query splitter, job dispatcher, task manager, result collector and result formatter. They enable us to develop a cooperative GRIDBLAST system between a server and heterogeneous remote worker nodes: which consist of various computer
architectures, different BLAST implementations and different Job schedulers operated by local resource management policy. The OBIGbs can execute 29,941 PSI-BLAST query sequences in 8.31 hours when using 230 CPUs in total and can return a 1.37 Giga byte result file.

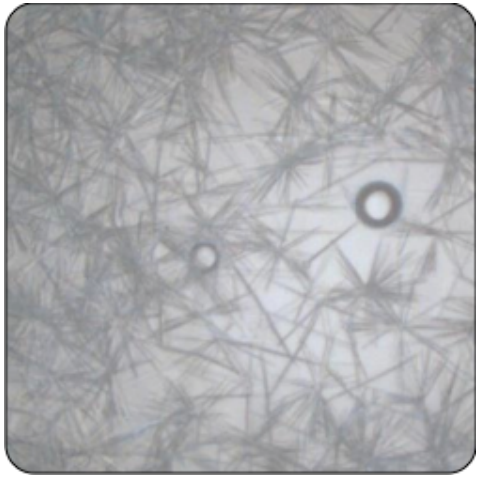




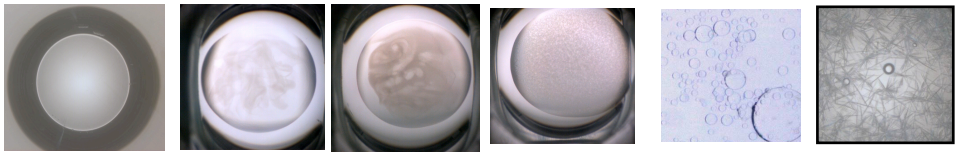
FPGA

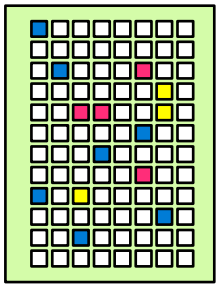
A method is described for enumerating the frequencies of DNA subsequences on a system comprising a host computer and a field programmable gate array (FPGA) board with one FPGA. Frequencies of subsequences with lengths of up to *K*0 + *K*1 + *K*2 (24 in the current implementation) are enumerated in three phases. In these three phases, subsequences with lengths of up to *K*0, *K*0 +*K*1, and *K*0 +*K*1 +*K*2, respectively, are enumerated; these three phases are executed simultaneously on a pipelined circuit, resulting in high performance.
The enumeration of frequent subsequences in databases, which are becoming larger and larger, will enable subsequences that are unique and/or repeatedly used in many parts of the sequences to be found.





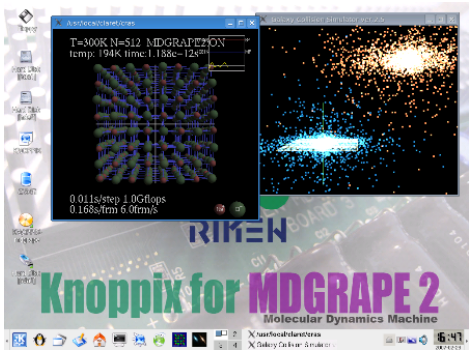
結晶スクリーニング結果に基づく、タンパク質結晶化条件用SVMモデルアレイの作成





Instant Computing

簡便に高度なシステムインテグレーションを提供するために開発



タンパク質配列情報の網羅的情報検索システムのInterProScanのクラスターによるクラスター計算機実行環境をインスタントに構築することができるライブCD。

分子シミュレーションに関するソフトウェアを収録し、クラスター計算機実行環境を構築できるようにしたもの。

大規模な計算機センターを単一イメージで構築することができるようにしたライブCD(Condor, Gfarm,PVFS, Gangulia等が構築可能）

最新のリマスタリングイメージは、KNOPPIX for CUDA

KNOPPIXとは

- KNOPPIX（クノーピクス）とは、CD-ROMまたはDVD-ROMから起動することが可能なDebianベースのLinuxディストリビューション。
- ドイツのKlaus KnopperがDebianパッケージを元に開発しており、日本語版は独立行政法人産業技術総合研究所が日本語化をはじめとする、日本の国情にあわせた様々な機能を追加して配布を行っている。 (Wikipedia)

Ex. Knoppix for Math, Knoppix for Edu etc..

最初の試みは

- バイオインフォマティックスの専門的なアプリケーションをインストールしたリマスタリングしたKnoppix for Bioに触発されて、より高度な計算機の利用ができるHigh Throughput Computing Editionを開発。

収録アプリケーションにバイオインフォマティクス関連ソフトウェア（BLAST, HMMER等）を収録＋Condor+PVFS

KNOB HTC Editionとは

KNOBとは

Knoppix for Bio - Bioにカスタマイズしたknoppix

http://knob.sourceforge.jp/

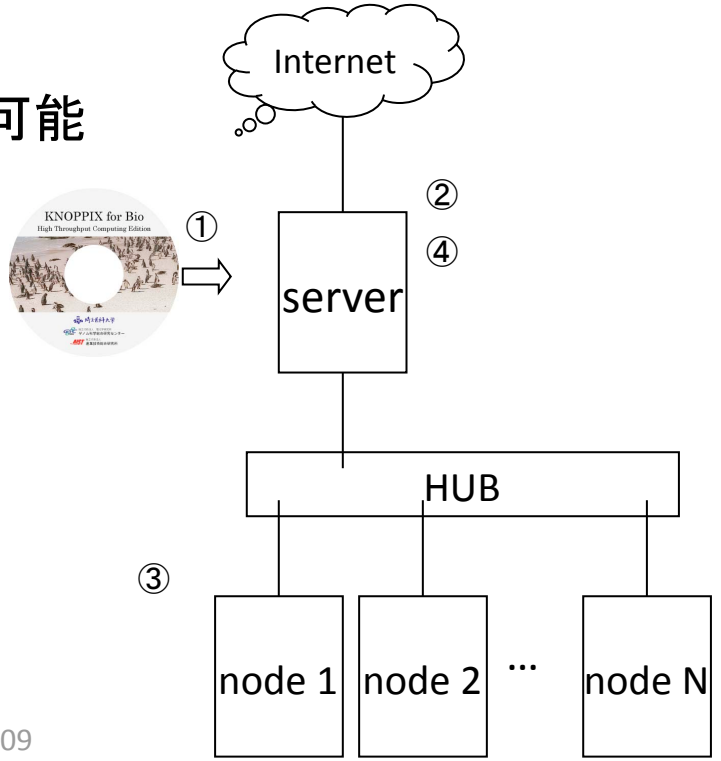
KNOBとの違い

●Condor によるジョブスケジューリングが可能

●PVFS2による並列ファイルシステムの構築が可能

KNOB HTC Edition 起動イメージ

1. serverをCDで起動
2. serverでネットブートのための設定
3. nodeがネットブート
4. serverでCondor, PVFS2等の設定
5. 完了



Condor とは

●米国ウィスコンシン大学マディソン校におけるCondor Research Project により開発・配布されているフリーのジョブスケジューラ

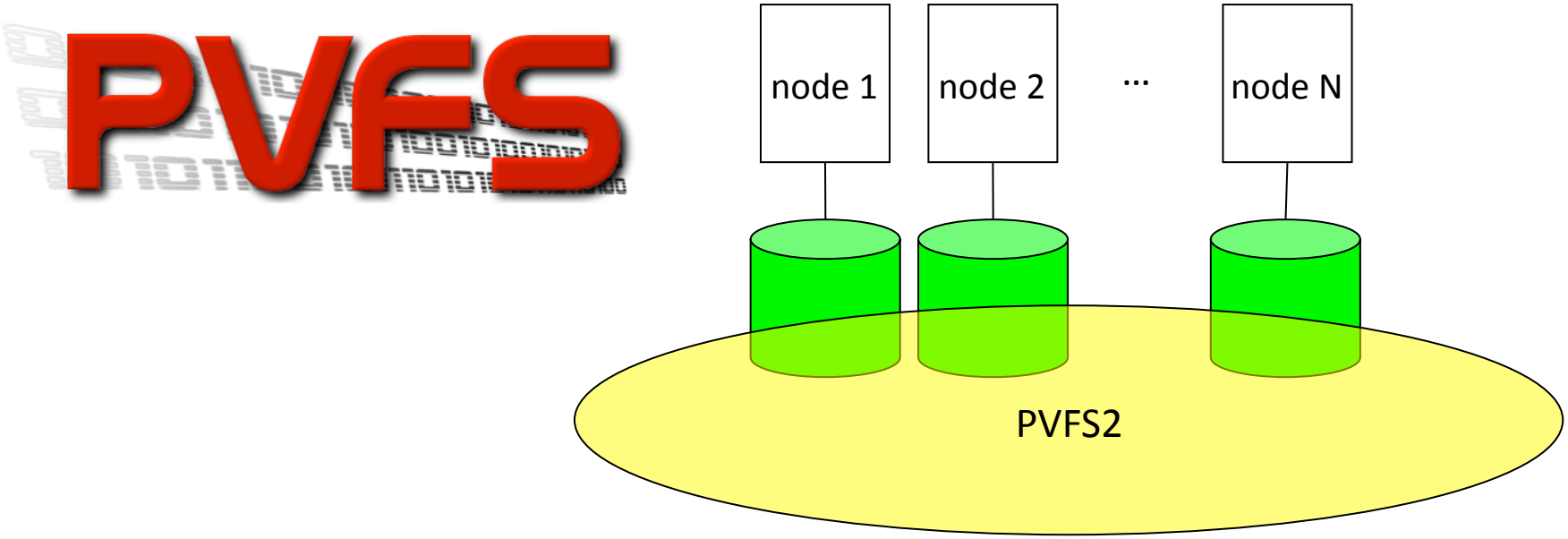
●複数のノードから構成されるCondor Poolと呼ばれる計算資源にジョブを効率よく分散して処理するシステム

Condor Project Homepage

http://www.cs.wisc.edu/condor/

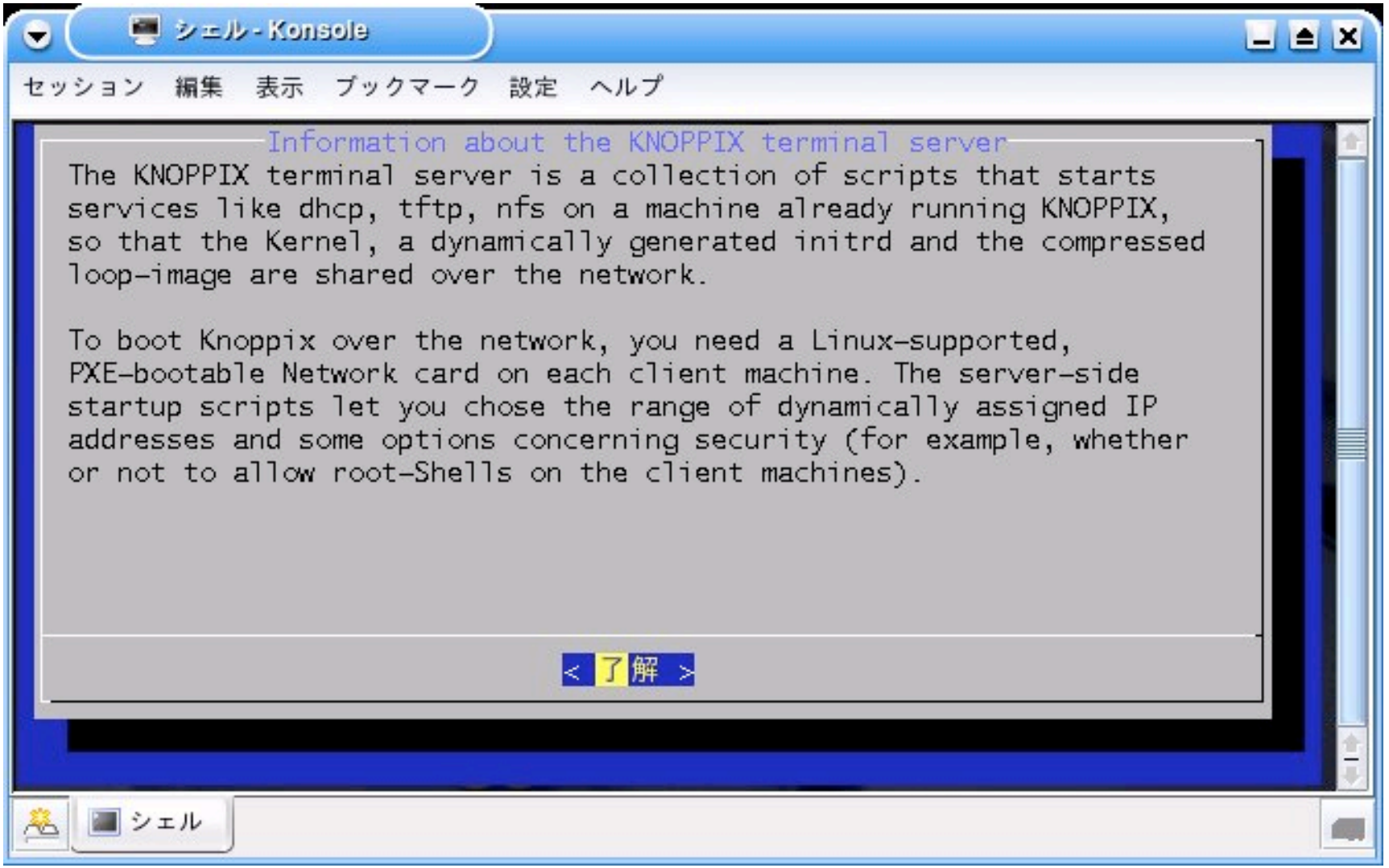
PVFS2 (Parallel File System 2)とは

- 並列ファイルシステムの一種
- 複数のノードがそれぞれディスクスペースを提供することによって，1つの巨大なディスクスペースを構築



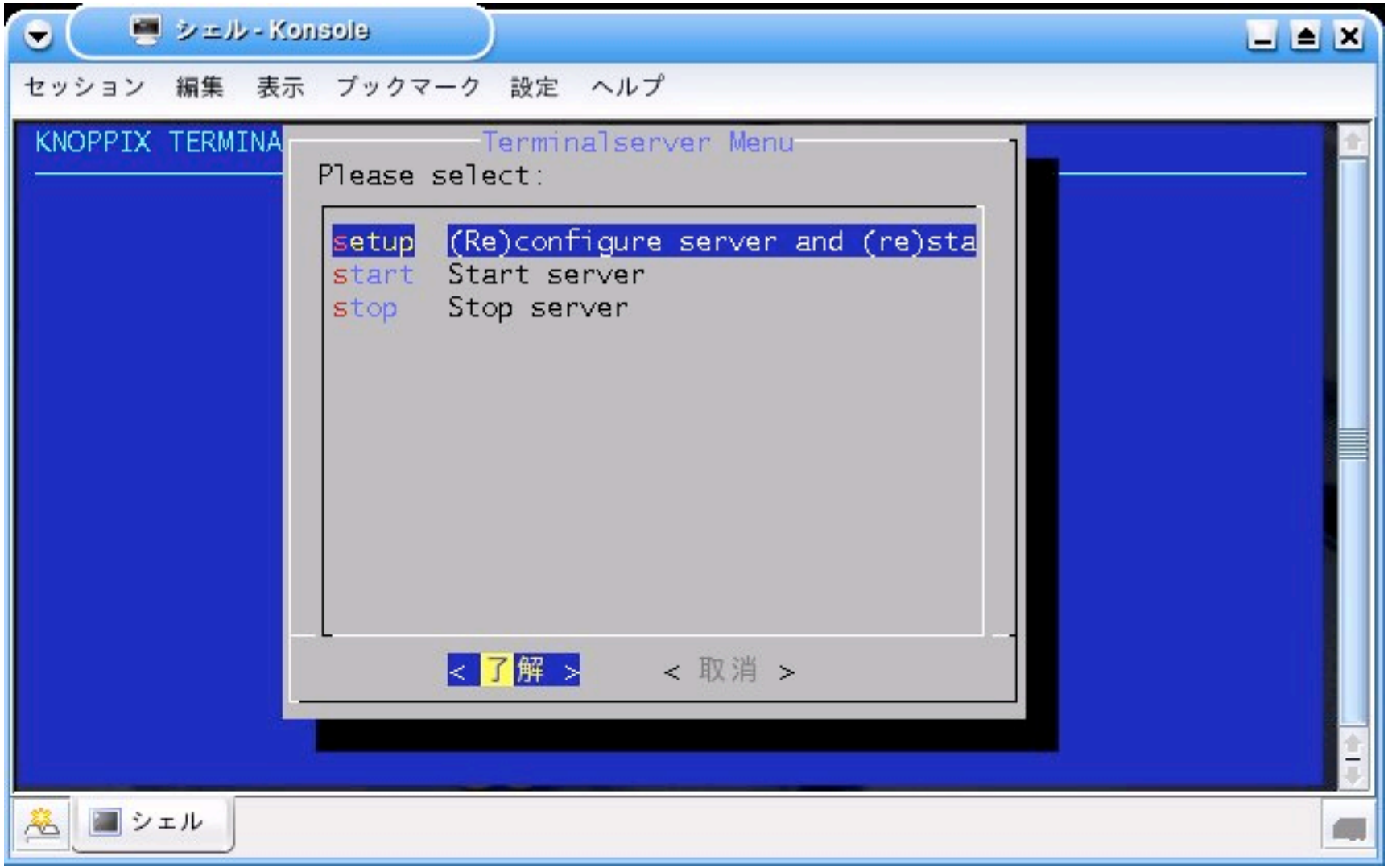
The Parallel File System Project

http://www.pvfs.org/pvfs2/

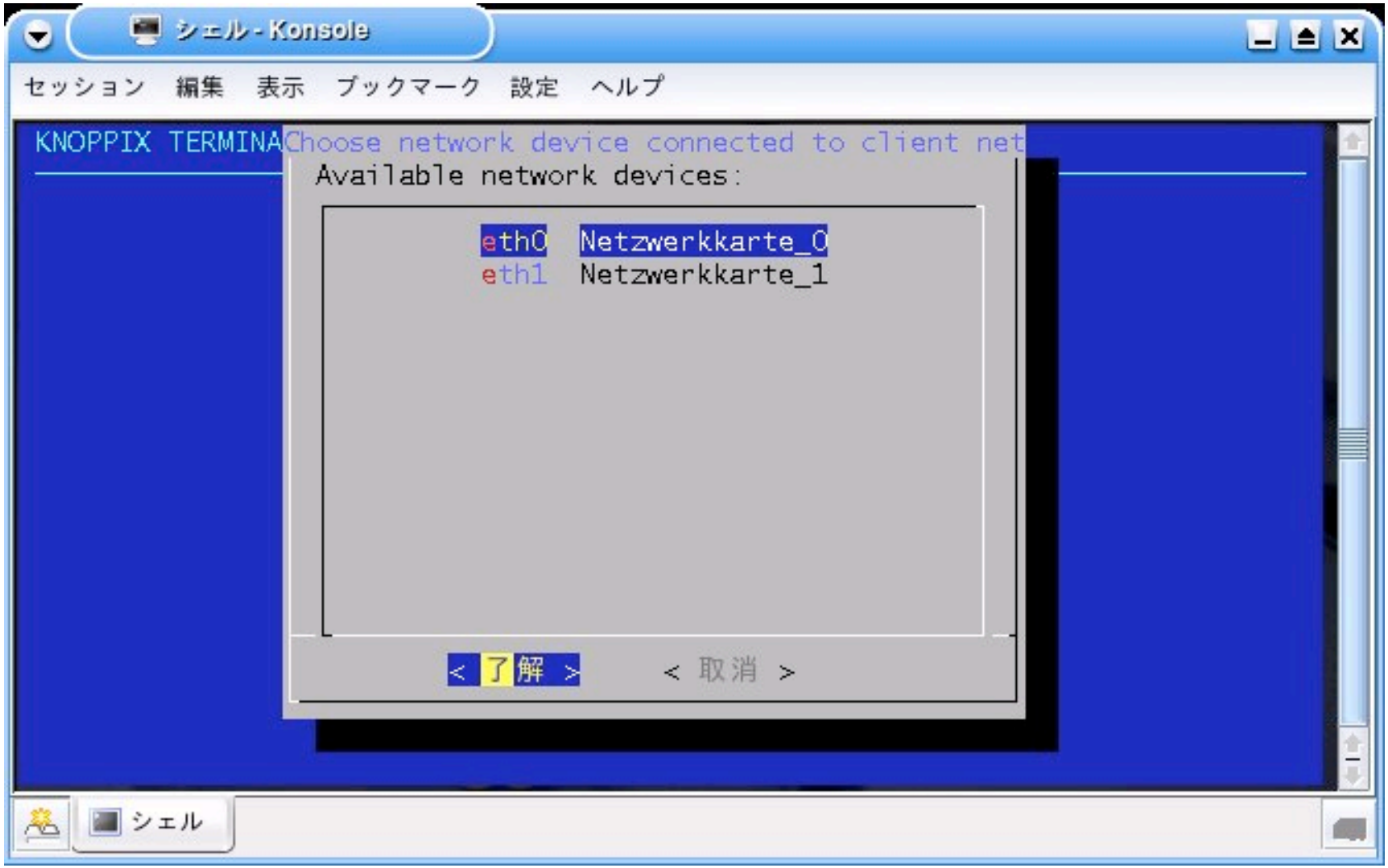


knoppix-terminalserverが起動します。<了解>を押してください。

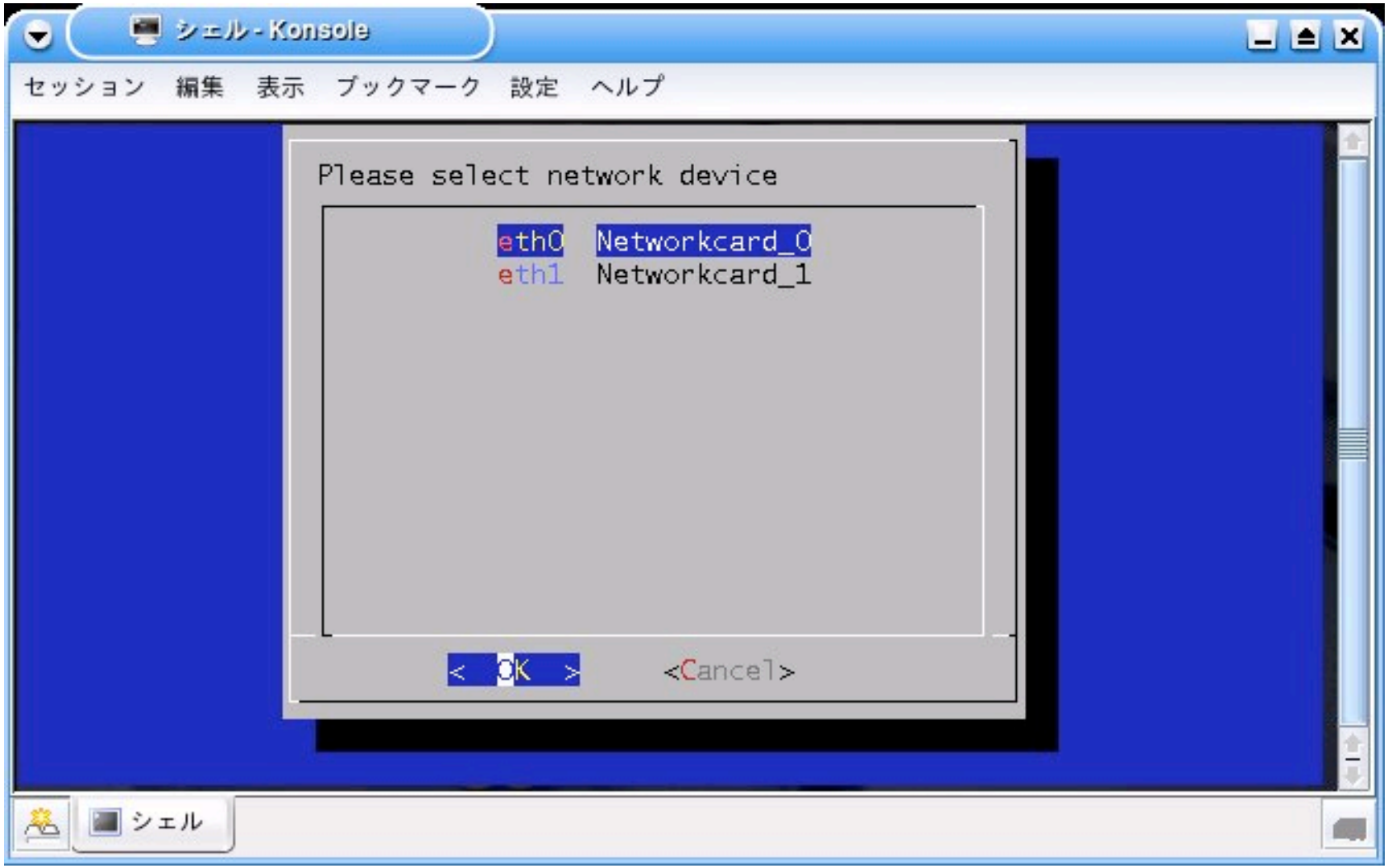
knoppix-terminalserverはネットワークブートのため，DHCPやTFTPの設定をするためのコマンドです。



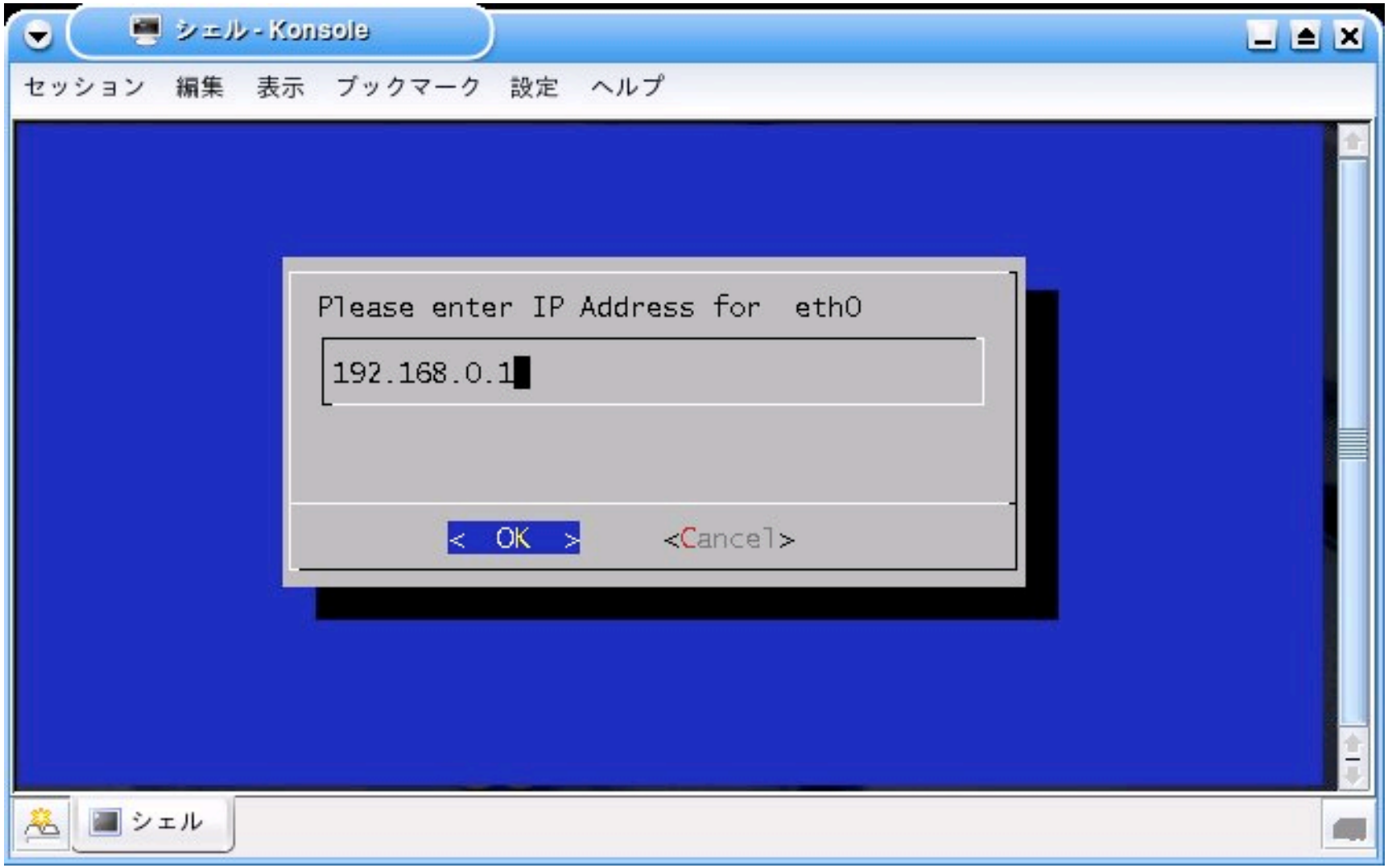
MENUからsetupを選択し，<了解>を押してください。



DHCPのサービスを起動するデバイスを選択して<了解>を押してください。



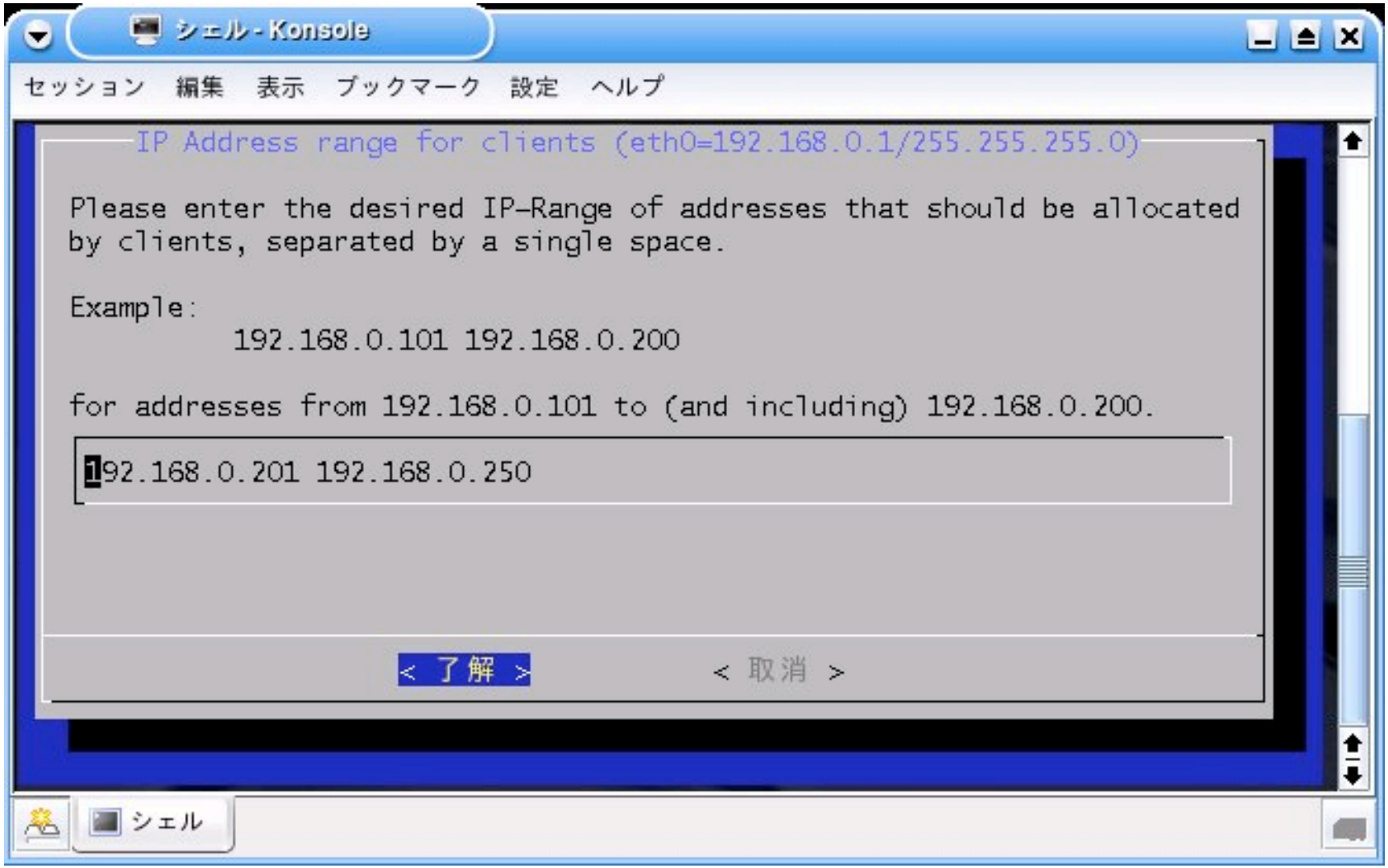
先程選択したデバイスを選択し，< OK >を押してください。



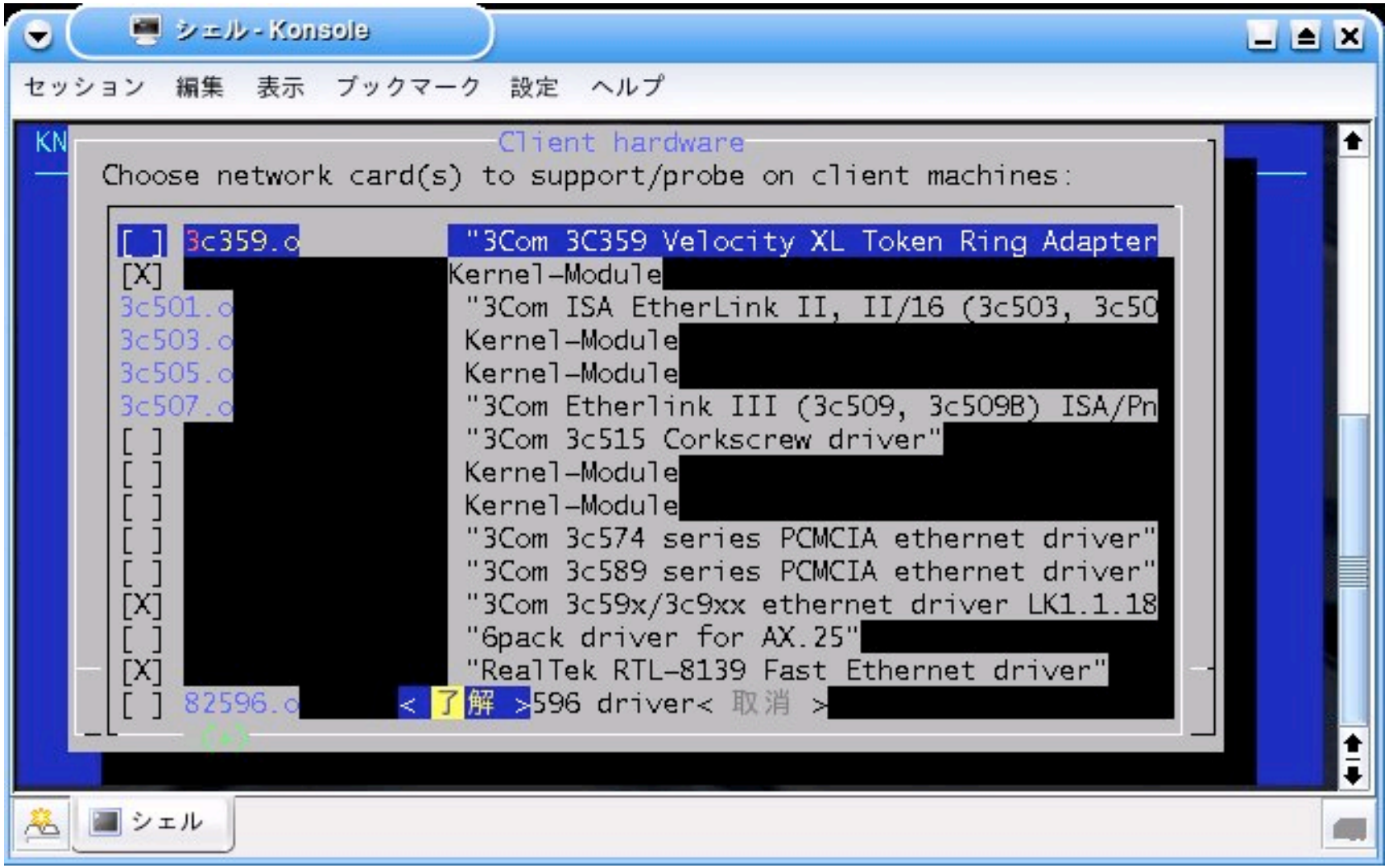
IPアドレス，ネットマスク，ブロードキャスト，ゲートウェイ，DNSサーバの項目を入力し，< OK >を押してください。

注意：DNSサーバが存在しない場合，DNSサーバの項目には何も入力しないでください。

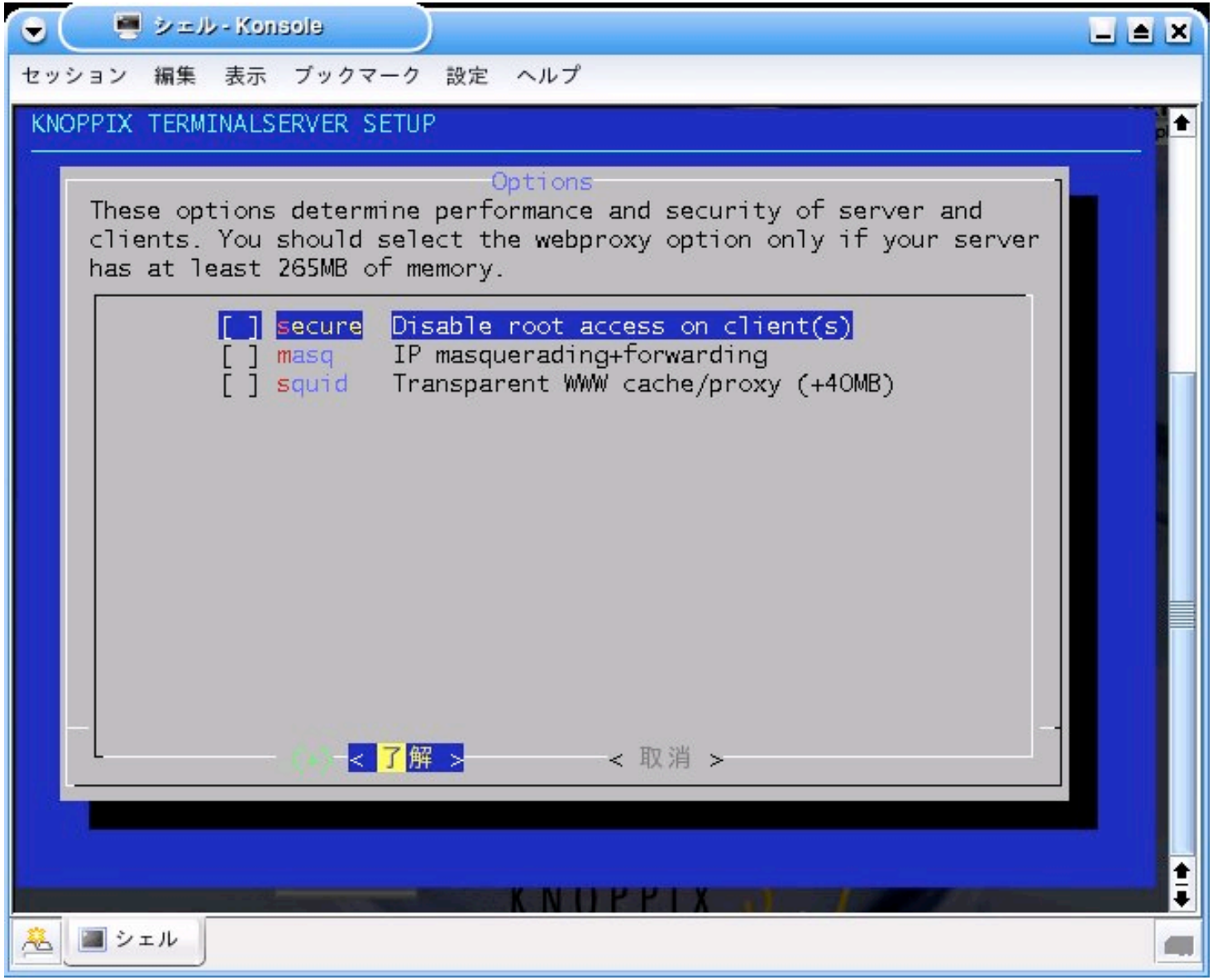
起動時にTimeout待ちが発生し、起動が遅くなります。



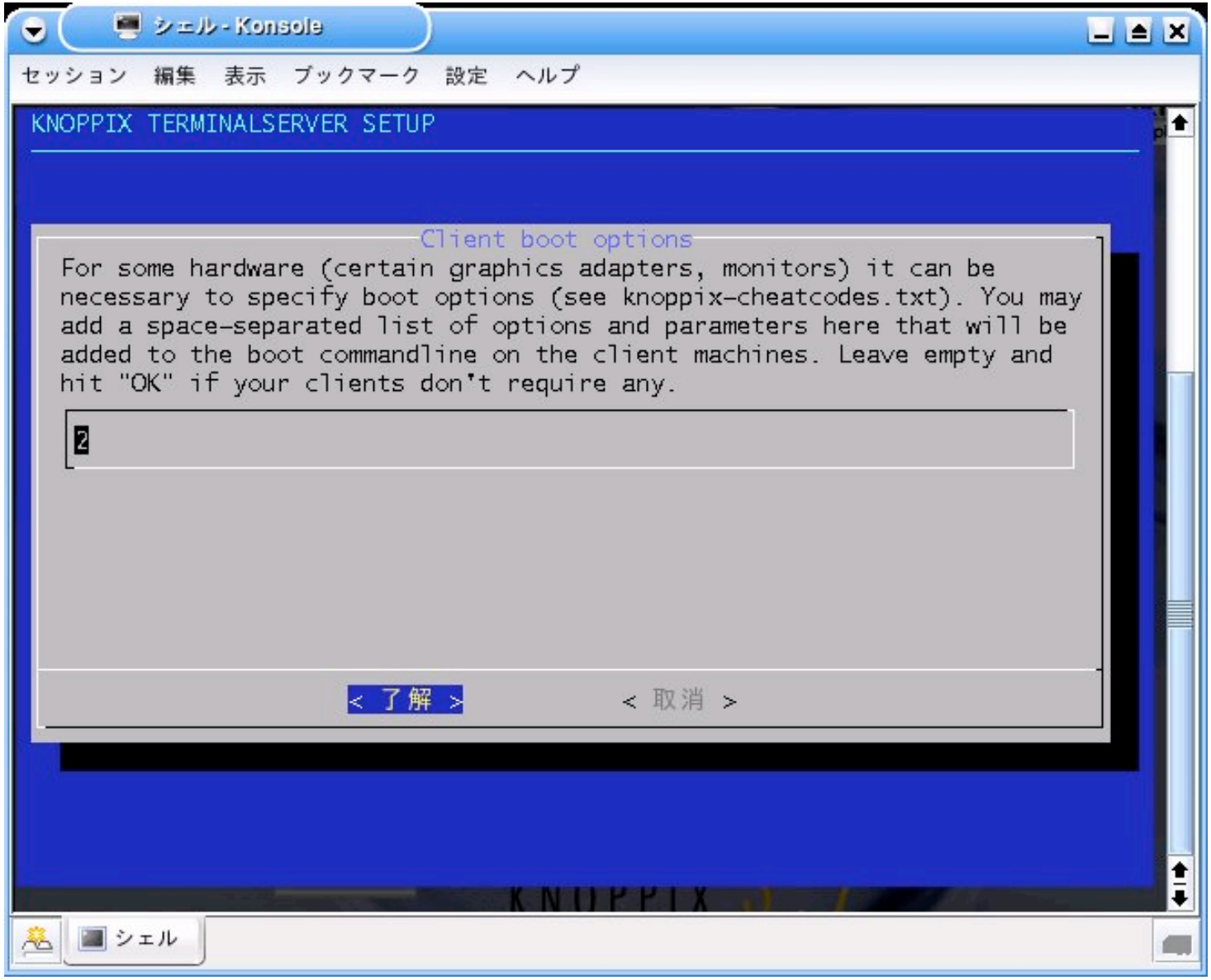
DHCPでクライアントに割り当てるIPの範囲を指定します。



クライアントのネットワークデバイスのドライバを選択します。



オプションを選択します。そのまま<了解>を押してください。



クライアントの起動オプションを入力します。デフォルトでは2（テキストモードで起動）が入力されています。

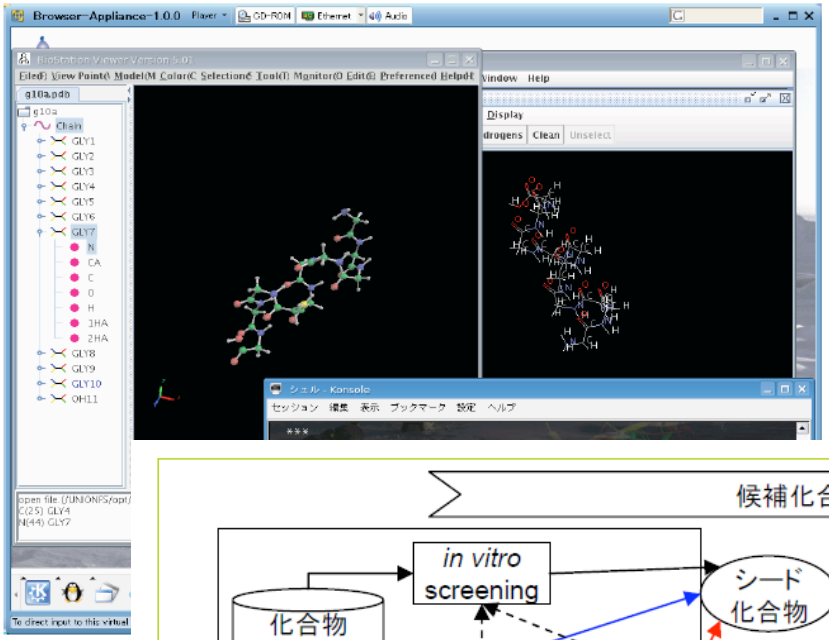
我々が学んだこと

- ローカルストレージを使わなくても並列ファイルシステムを利用することで、仮想的なディスクを利用することができる。
- クラスター計算機が、CD1枚から構築することがきる。
- スパコンの限界の縮図としてシステムを評価することができた。
  - 例：並列ファイルシステムのメタデータ自体の容量に起因する問題や、ディスクスペースの枯渇時に起こる挙動等。

次の試み

- よりアプリケーションを特定して、目的別にCDイメージを用意することで、ユーザに簡便に環境構築をすることで、利用してもらう。
- システムの設定手順をGUI的（ブロードバンドルータの設定のような形）にすることで、簡便に構築させる。

分子シミュレーションをターゲットにする



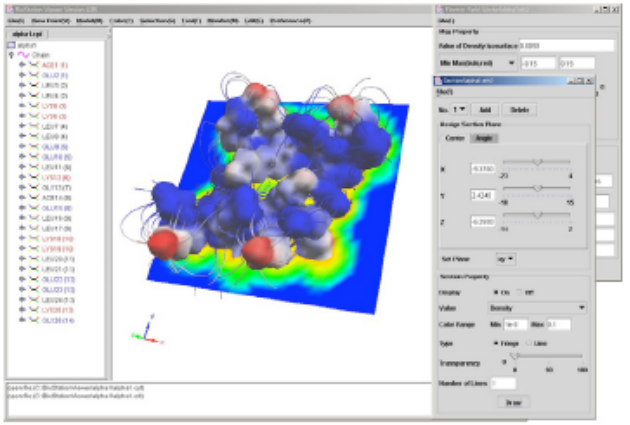


図3　電場と電子密度
分子の等電子密度面の静電ポテンシャルの値により色分けし、分子周囲の電場を電気力線で表現している。また、任意の断面上で、静電ポテンシャルや電子

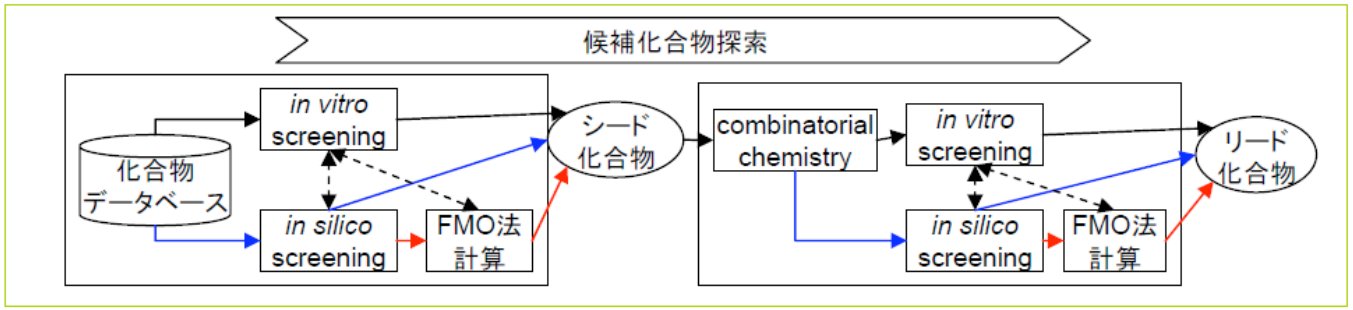


図2　候補化合物探索におけるAdvance/BioStationによるFMO計算の活用イメージ

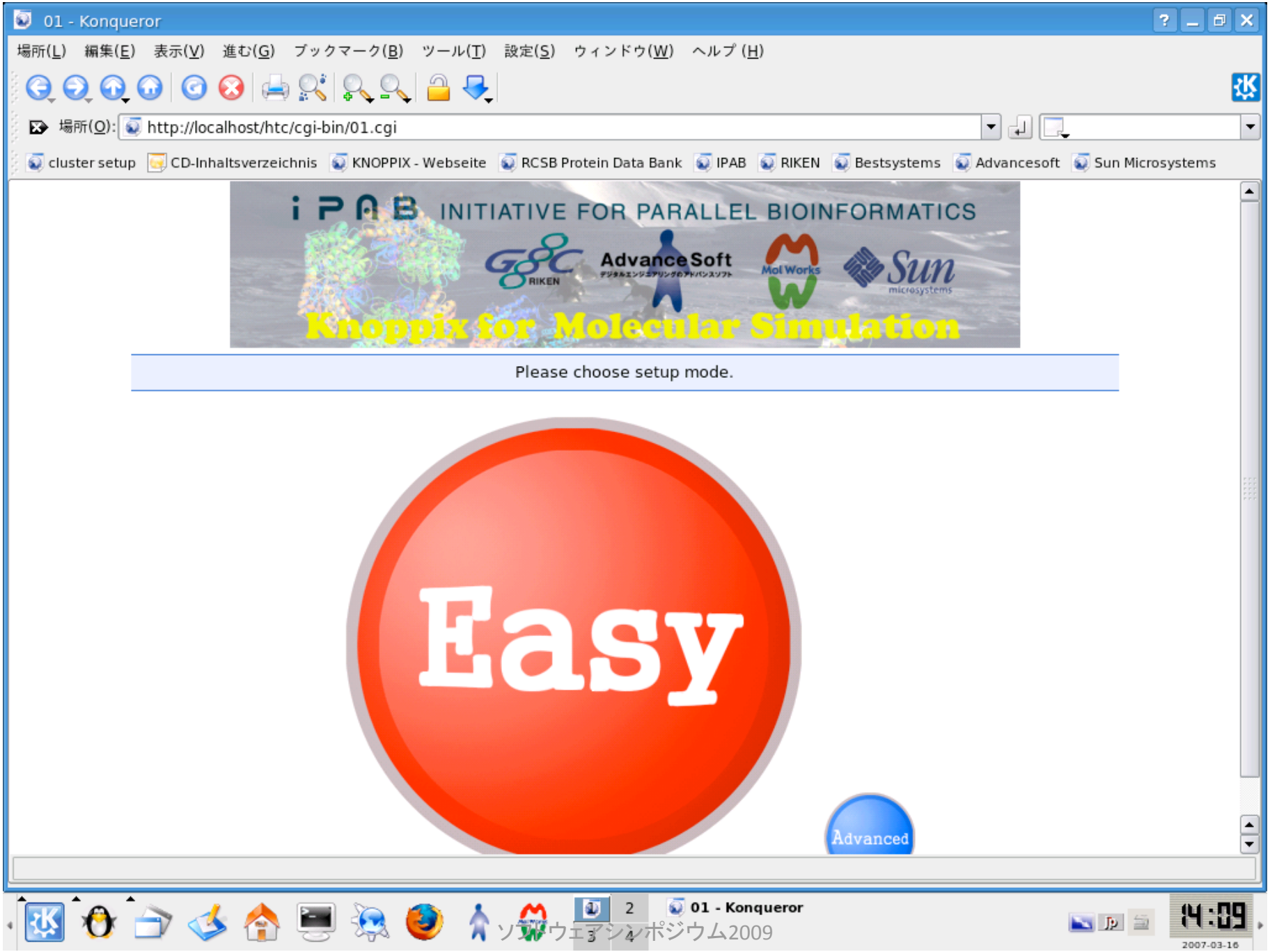
どんな計算機で動かすのか

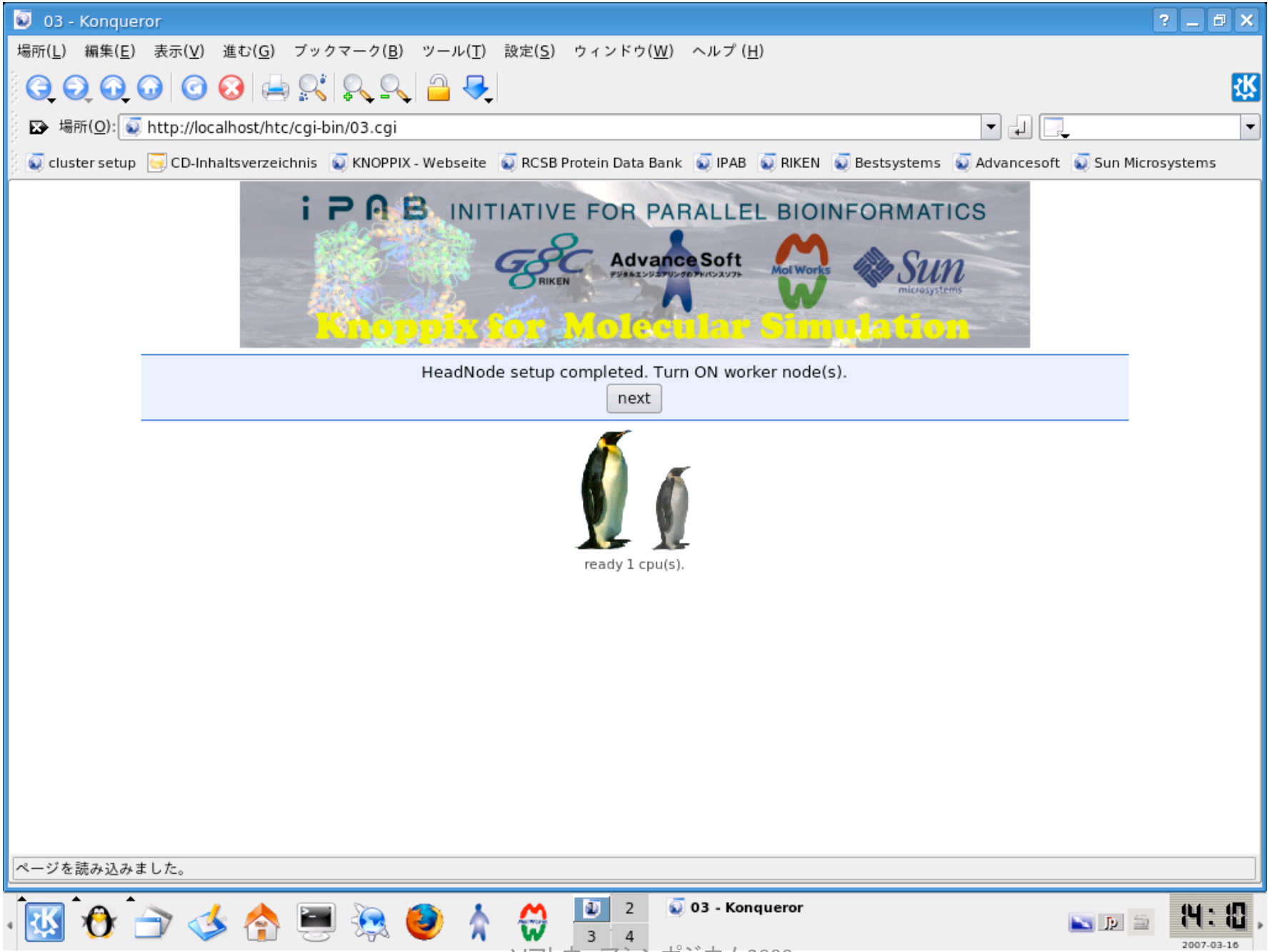
- メモリは、1 Core 当たり、 1GBytesを想定
- マルチコア計算機
  - 1ノード、1Core
  - 1ノード、マルチコア （2 < n < 8)
- クラスターノード
  - マルチノード、1Core (m>2)
  - マルチノード、マルチコア （ 1 < n < 8, 1 < m < 32くらい）

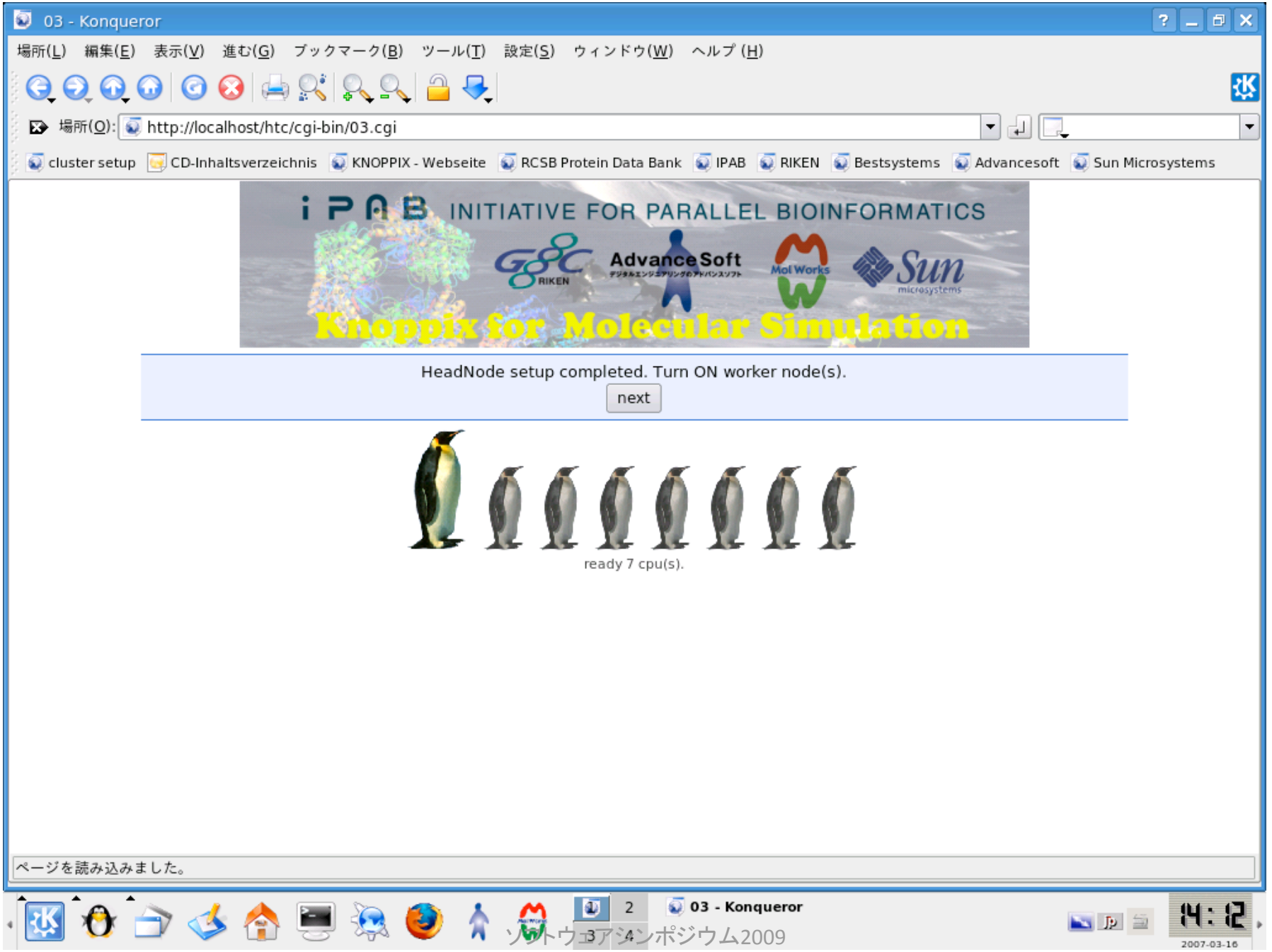
なにがはいっているのか？

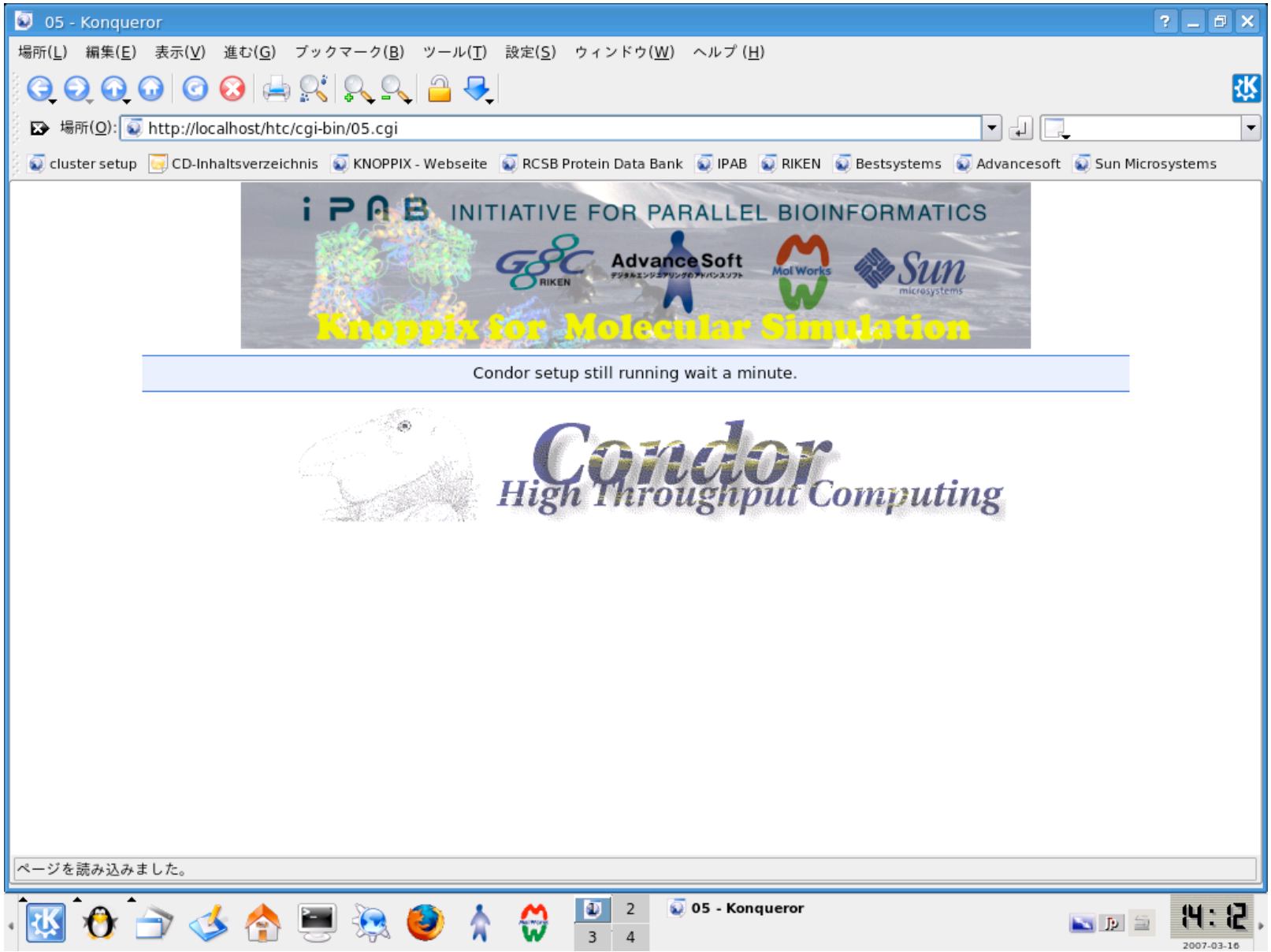
- 起動システム
- Knoppix 5.0.1 Linux Kernel 2.6.17 Japanese edition
- 動作環境
- プロセッサー Pentium4 2.0 GHz 以上推奨 ,メモリ/プロセッサー 512Mバイト以上、1Gバイト以上推奨, ネットワーク 100BASE-TX以上
- Knoppix for Molecure Simulationに組み込まれたソフトウェア。
- General library tool
- ・bioperl 1.4-1
- Sequence Search
- ・blast2 2.2.15.20061015-1
- ・ncbi-epcr 1.2.0-3
- ・ncbi-tools-(x11|bin) 6.1.20061015-1
- ・fastlink 4.1P-fix92-1
- Viewer
- ・BioStation Viewer※
- ・rasmol 2.7.2.1.1-4
- ・seaview 20060918-1
- ・xmakemol-gl 5.15-1
- ・mozilla-biofox 1.1.2+0-1
- ・xbs 0-7.3
- ・gdis 0.86-2
- Drawer(editor)
- ・MolWorks 2.0
- ・easychem 0.6.2
- ・chemtool 1.6.9-1
- Simulation
- ・abinitmp demo_ver.2.0
- ・reduce 2.21
- ・polyxmass 0.9.2
- ・psi3 3.2.3-1
- ・gromacs 3.3.1-2
- ・primer3, perlprimer 1.0b-1, 1.1.13-1
- converter
- ・babel 1.6

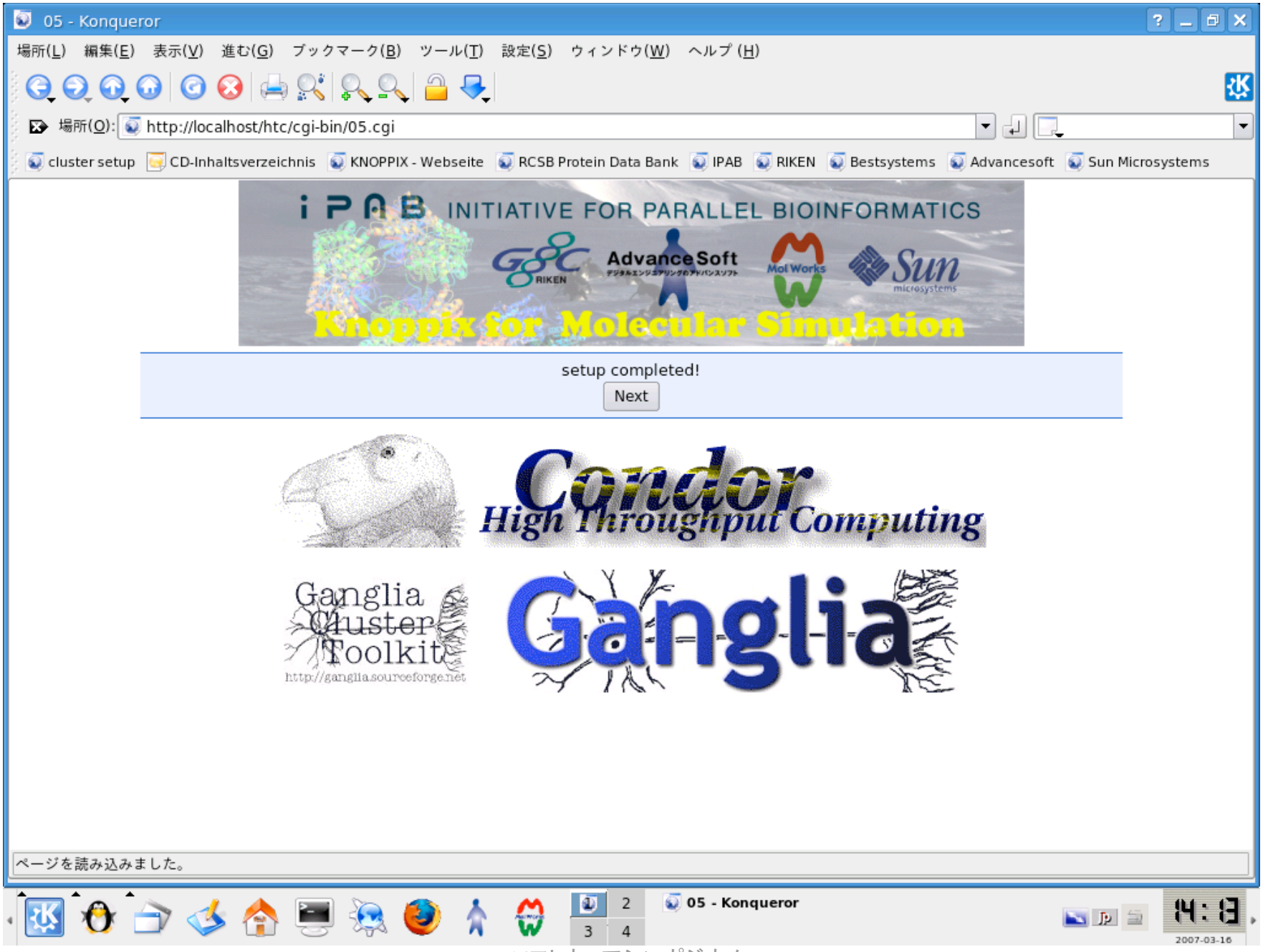
- MPI
- Condor

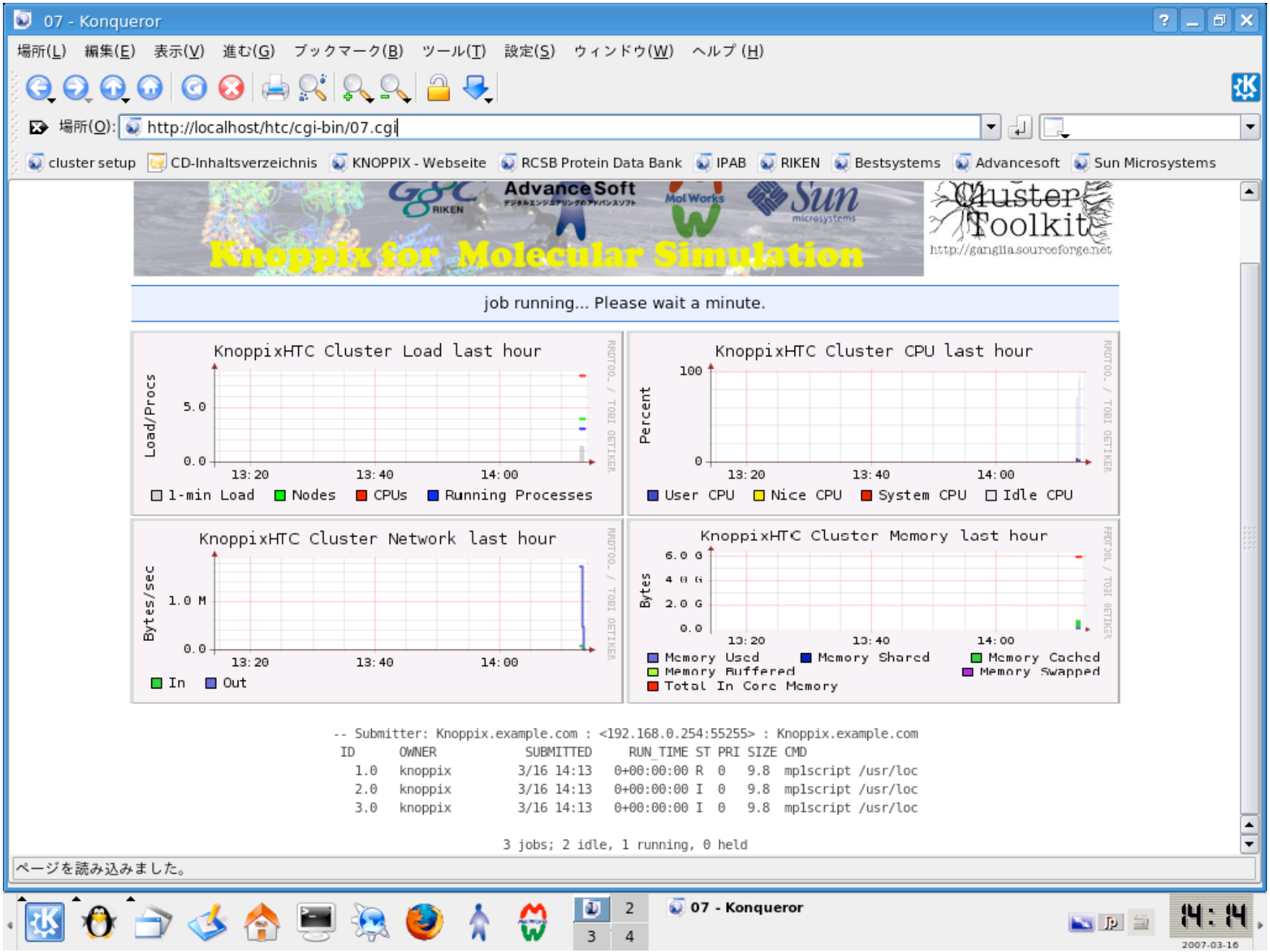

01 - Konqueror
場所(L) 編集(E) 表示(V) 進む(G) ブックマーク(B) ツール(T) 設定(S) ウィンドウ(W) ヘルプ (H)
場所(O): http://localhost/htc/cgi-bin/01.cgi
cluster setup
CD-Inhaltsverzeichnis
KNOPPIX - Webseite
RCSB Protein Data Bank
IPAB
RIKEN
Bestsystems
Advancesoft
Sun Microsystems
iPAB INITIATIVE FOR PARALLEL BIOINFORMATICS
Knoppix for Molecular Simulation
Please choose setup mode.
Easy
Advanced
01 - Konqueror
14:09
2007-03-16


03 - Konqueror
場所(L) 編集(E) 表示(V) 進む(G) ブックマーク(B) ツール(T) 設定(S) ウィンドウ(W) ヘルプ (H)
場所(O): http://localhost/htc/cgi-bin/03.cgi
cluster setup
CD-Inhaltsverzeichnis
KNOPPIX - Webseite
RCSB Protein Data Bank
IPAB
RIKEN
Bestsystems
Advancesoft
Sun Microsystems
iPAB INITIATIVE FOR PARALLEL BIOINFORMATICS
RIKEN
AdvanceSoft
Mol Works
Sun microsystems
Knoppix for Molecular Simulation
HeadNode setup completed. Turn ON worker node(s).
next
ready 1 cpu(s).
ページを読み込みました。
03 - Konqueror
2007-03-16


03 - Konqueror
場所(L) 編集(E) 表示(V) 進む(G) ブックマーク(B) ツール(T) 設定(S) ウィンドウ(W) ヘルプ (H)
場所(O): http://localhost/htc/cgi-bin/03.cgi
cluster setup
CD-Inhaltsverzeichnis
KNOPPIX - Webseite
RCSB Protein Data Bank
IPAB
RIKEN
Bestsystems
Advancesoft
Sun Microsystems
iPAB INITIATIVE FOR PARALLEL BIOINFORMATICS
RIKEN
AdvanceSoft
Mol Works
Sun microsystems
Knoppix for Molecular Simulation
HeadNode setup completed. Turn ON worker node(s).
next
ready 7 cpu(s).
ページを読み込みました。
03 - Konqueror
2007-03-16


05 - Konqueror
場所(L) 編集(E) 表示(V) 進む(G) ブックマーク(B) ツール(T) 設定(S) ウィンドウ(W) ヘルプ (H)
場所(O): http://localhost/htc/cgi-bin/05.cgi
cluster setup
CD-Inhaltsverzeichnis
KNOPPIX - Webseite
RCSB Protein Data Bank
IPAB
RIKEN
Bestsystems
Advancesoft
Sun Microsystems
iPAB
INITIATIVE FOR PARALLEL BIOINFORMATICS
RIKEN
AdvanceSoft
Mol Works
Sun
microsystems
Knoppix for Molecular Simulation
Condor setup still running wait a minute.
Condor
High Throughput Computing
ページを読み込みました。
05 - Konqueror
14:12
2007-03-16


05 - Konqueror
場所(L) 編集(E) 表示(V) 進む(G) ブックマーク(B) ツール(T) 設定(S) ウィンドウ(W) ヘルプ (H)
場所(O): http://localhost/htc/cgi-bin/05.cgi
cluster setup
CD-Inhaltsverzeichnis
KNOPPIX - Webseite
RCSB Protein Data Bank
IPAB
RIKEN
Bestsystems
Advancesoft
Sun Microsystems
iPAB INITIATIVE FOR PARALLEL BIOINFORMATICS
AdvanceSoft
Mol Works
Sun microsystems
Knoppix for Molecular Simulation
setup completed!
Next
Condor
High Throughput Computing
Ganglia
Cluster
Toolkit
http://ganglia.sourceforge.net
Ganglia
ページを読み込みました。
05 - Konqueror
2007-03-16


07 - Konqueror
場所(L) 編集(E) 表示(V) 進む(G) ブックマーク(B) ツール(T) 設定(S) ウィンドウ(W) ヘルプ(H)
場所(O): http://localhost/htc/cgi-bin/07.cgi
cluster setup
CD-Inhaltsverzeichnis
KNOPPIX - Webseite
RCSB Protein Data Bank
IPAB
RIKEN
Bestsystems
Advancesoft
Sun Microsystems
Knoppix for Molecular Simulation
Cluster Toolkit
http://ganglia.sourceforge.net
job running... Please wait a minute.
KnoppixHTC Cluster Load last hour
Load/Procs
1-min Load
Nodes
CPUs
Running Processes
KnoppixHTC Cluster CPU last hour
Percent
User CPU
Nice CPU
System CPU
Idle CPU
KnoppixHTC Cluster Network last hour
Bytes/sec
In
Out
KnoppixHTC Cluster Memory last hour
Bytes
Memory Used
Memory Shared
Memory Cached
Memory Buffered
Memory Swapped
Total In Core Memory
-- Submitter: Knoppix.example.com : <192.168.0.254:55255> : Knoppix.example.com
ID OWNER SUBMITTED RUN_TIME ST PRI SIZE CMD
1.0 knoppix 3/16 14:13 0+00:00:00 R 0 9.8 mplscript /usr/loc
2.0 knoppix 3/16 14:13 0+00:00:00 I 0 9.8 mplscript /usr/loc
3.0 knoppix 3/16 14:13 0+00:00:00 I 0 9.8 mplscript /usr/loc
3 jobs; 2 idle, 1 running, 0 held
ページを読み込みました。
07 - Konqueror
14:14
2007-03-16

07 - Konqueror

場所(L) 編集(E) 表示(V) 進む(G) ブックマーク(B) ツール(T) 設定(S) ウィンドウ(W) ヘルプ(H)

場所(O): http://localhost/htc/cgi-bin/07.cgi

cluster setup | CD-Inhaltsverzeichnis | KNOPPIX - Webseite | RCSB Protein Data Bank | IPAB | RIKEN | Bestsystems | Advancesoft | Sun Microsystems

Knoppix for Molecular Simulation

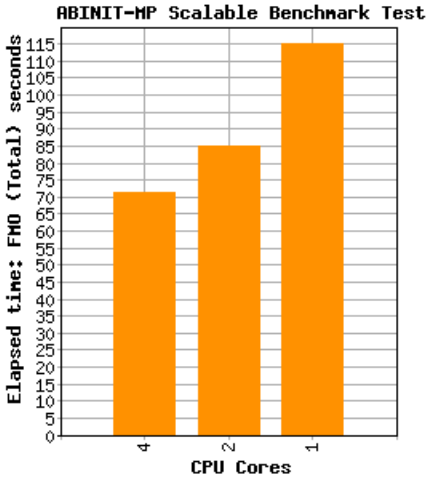
http://ganglia.sourceforge.net

test job completed!! result

You would like to perform a benchmark job, click "benchmark" button.

finish

benchmark



CPUs	Monomer SCF	Dimer SCF	FMO (Total)
4	48.1	23.3	71.4
2	54.6	30.3	84.9
1	75.2	39.9	115.1

ページを読み込みました。

07 - Konqueror

14:16

2007-03-16

関連システム（1）

HTTP-FUSEを使ったイメージ

アプリケーションを収録しているイメージを外部HTTPを使ったCLOOPファイルの取得により、分離したシステム

アプリケーションのバージョンアップに関する影響を最小限にした。

関連システム（2）

InterProScanを実装

6Gバイトのデータベース＋600Mバイトのメタデータ

本格的な専用環境のインスタントコンピューティングシステム

InterProScan4.1 Adaptation

Preprocessing

Sequence

Chunk size Turing

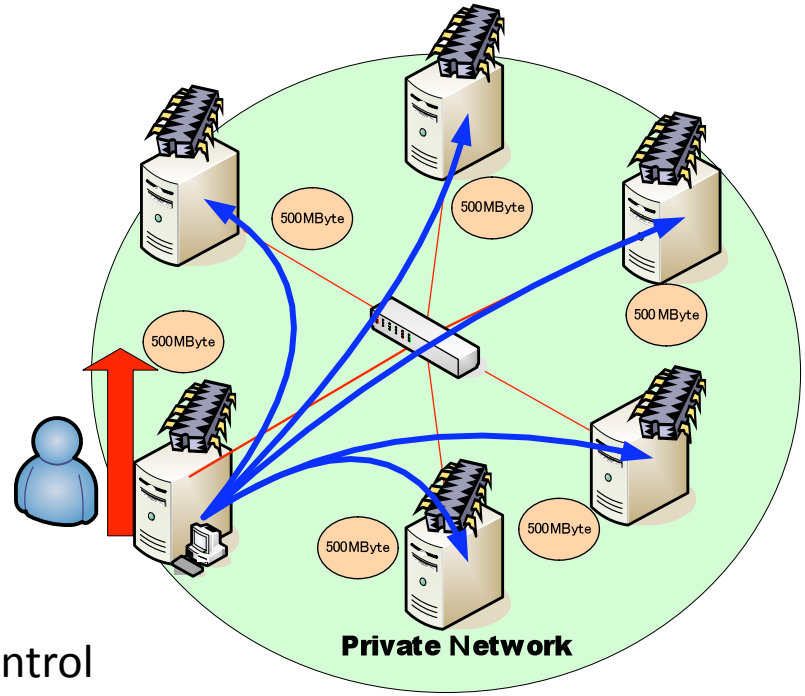
jobdispatch

Job status check loop

Condor adapter

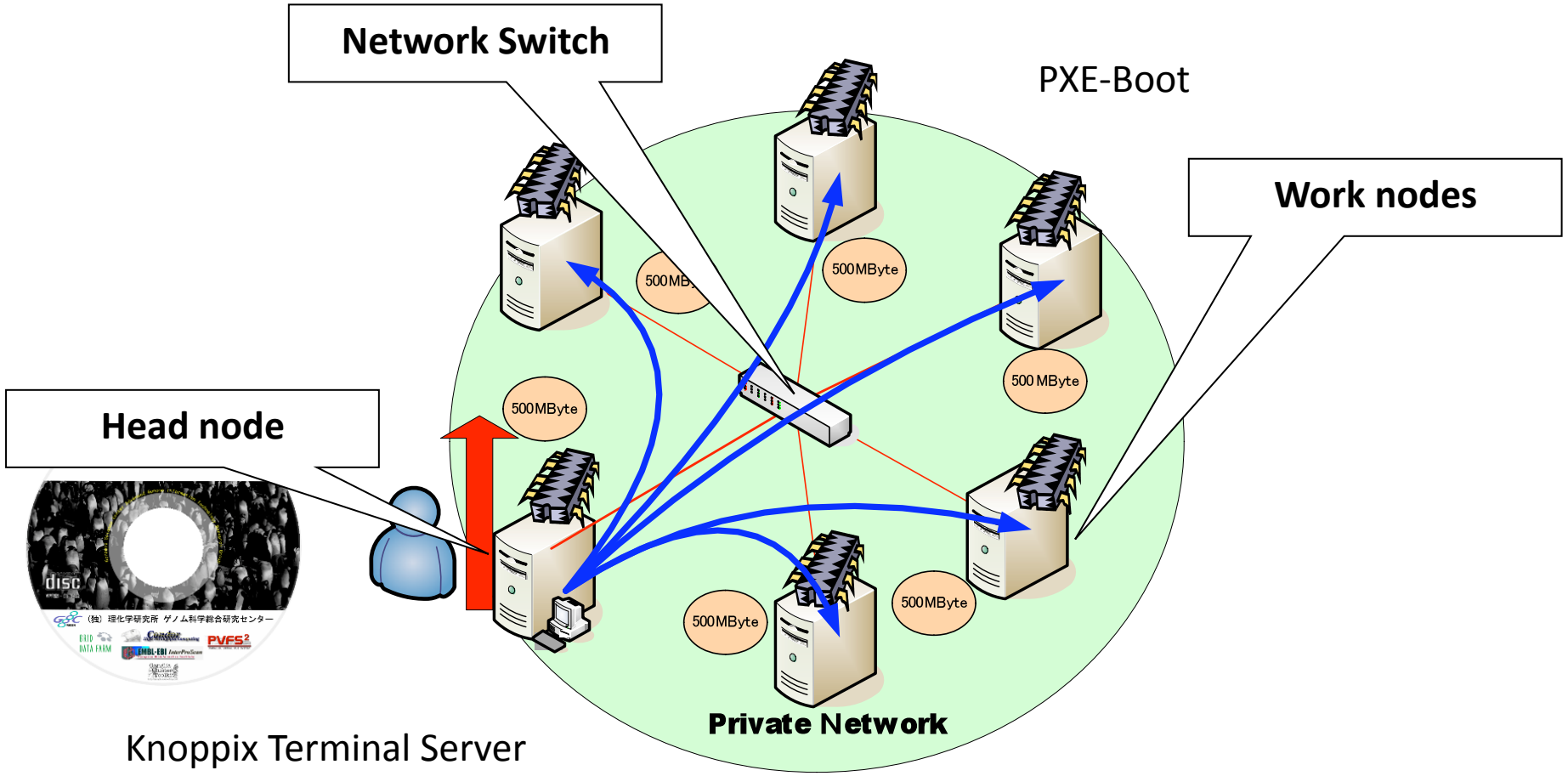
Postprocessing

Result Converter



Web GUI control

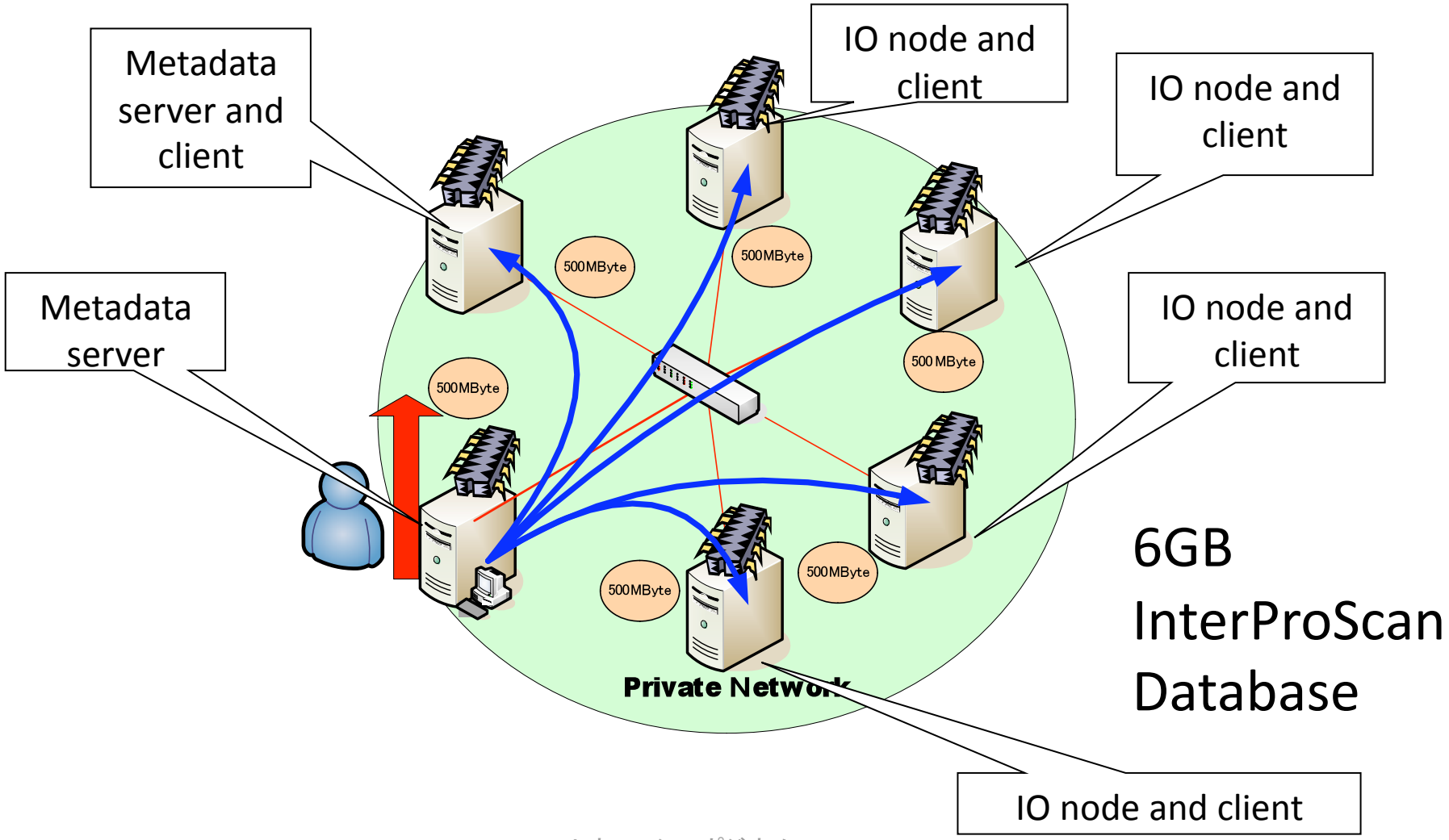
Deploying a Boot image with knoppix terminal server



Knoppix Terminal Server
DHCP, NFS

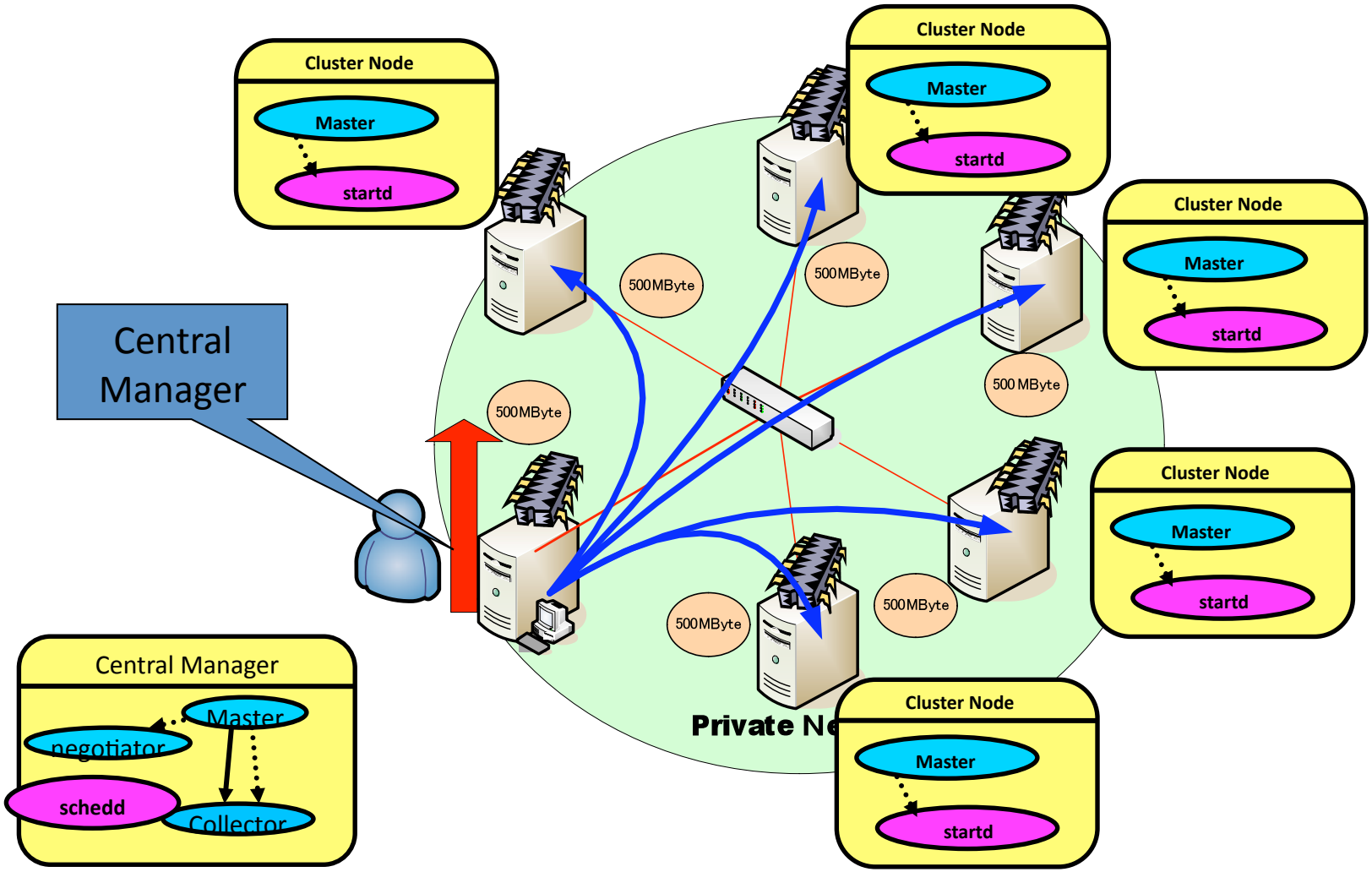
Each work node mounts CD-Rom storage by NFS on head node.

Creation of Scratch space by a parallel file system



6GB
InterProScan
Database

Pool Layout by High Throughput Computing Scheduler



我々がたどり着いたコンセプト

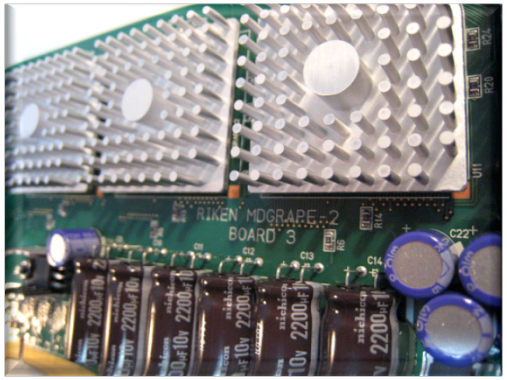
- 元の状態に戻せれば、他人は計算機を貸してくれそう。
- なんでもできるCDをつくるより、専門的なリマスタリングイメージを作成する方が喜ばれる。
- ソフトウェア開発も環境が固定されていれば、Debugしやすい。

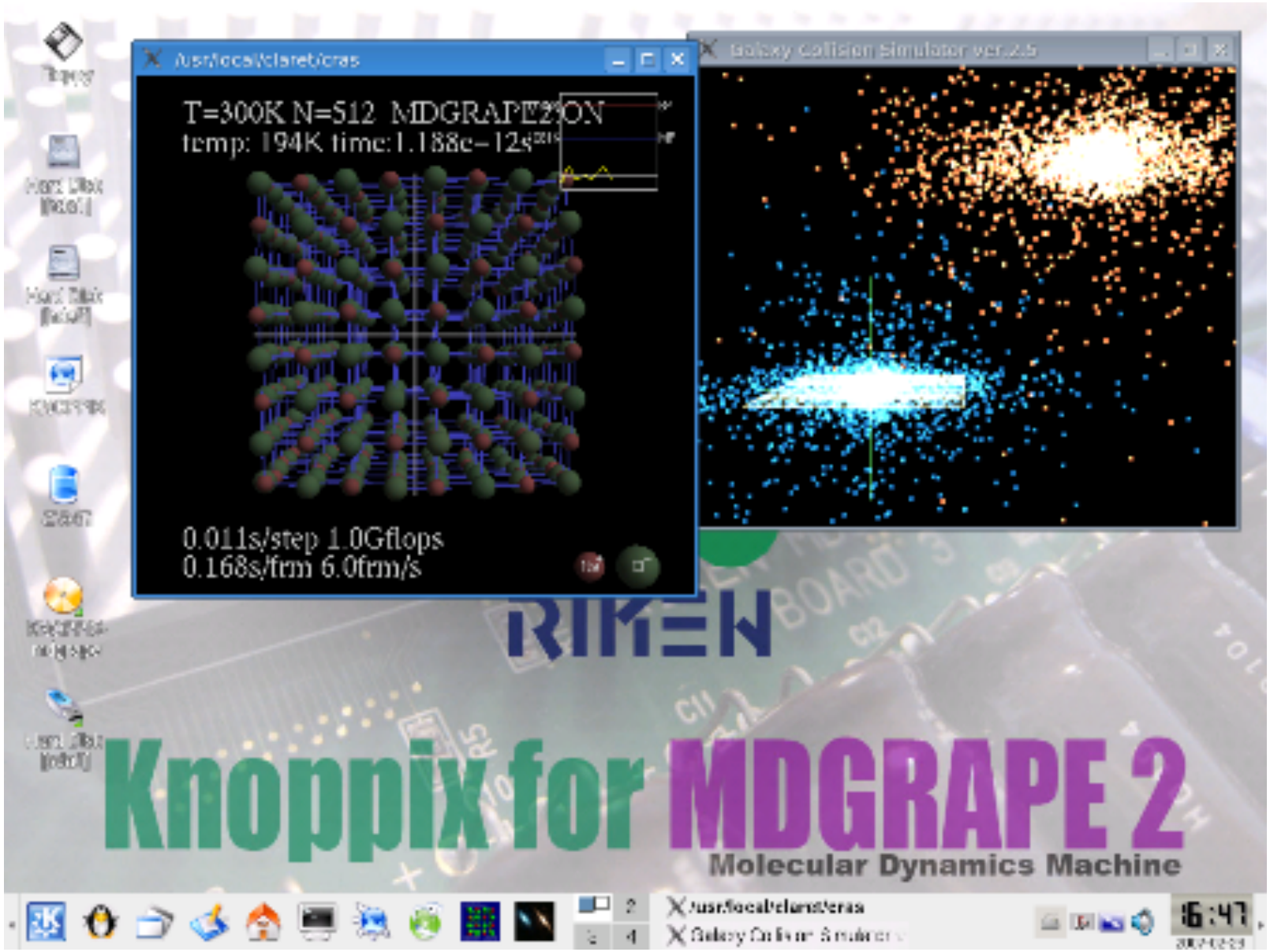
インスタントな環境を提供する市場の開拓

さらに別の可能性を模索する

- 専門的なソフトウェアの収集して、専用CDとして収録する。
  - →専門的なハードウェアを対象とした専用CDの開発。
- 理化学研究所で開発したMDGrapeの実行環境に特化したイメージを収録。

MDGrape専用ドライバー
N体問題プログラム
サンプルコード
SDK等




/usr/local/claret/cras
T=300K N=512 MDGRAPE2ON
temp: 194K time:1.188e-12s
0.011s/step 1.0Gflops
0.168s/frm 6.0frm/s
Galaxy Collision Simulator ver.2.5
RIKEN
Knoppix for MDGRAPE 2
Molecular Dynamics Machine

組織内部からの反応

- 専用ハードウェアのための環境をインスタントに構築するニーズは、潜在的に存在した。
  - ハードウェアの検証ツール（ベンチマークや、ストレスチェック）
  - 検証済み実行環境のスナップショット
    - 導入先のシステムでの諸問題（カーネル不一致、ライブラリ不足などなど）を回避することができる。

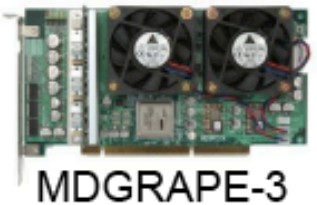
様々なアクセラレータに対応したい。

多様なアクセラレータ

使うのが簡単

- 専用計算機GRAPE、MDGRAPE
- FPGA
- GRAPE-DR
- ClearSpeed
- PLAYSTATION 3
- GPU

準汎用計算機


MDGRAPE-3

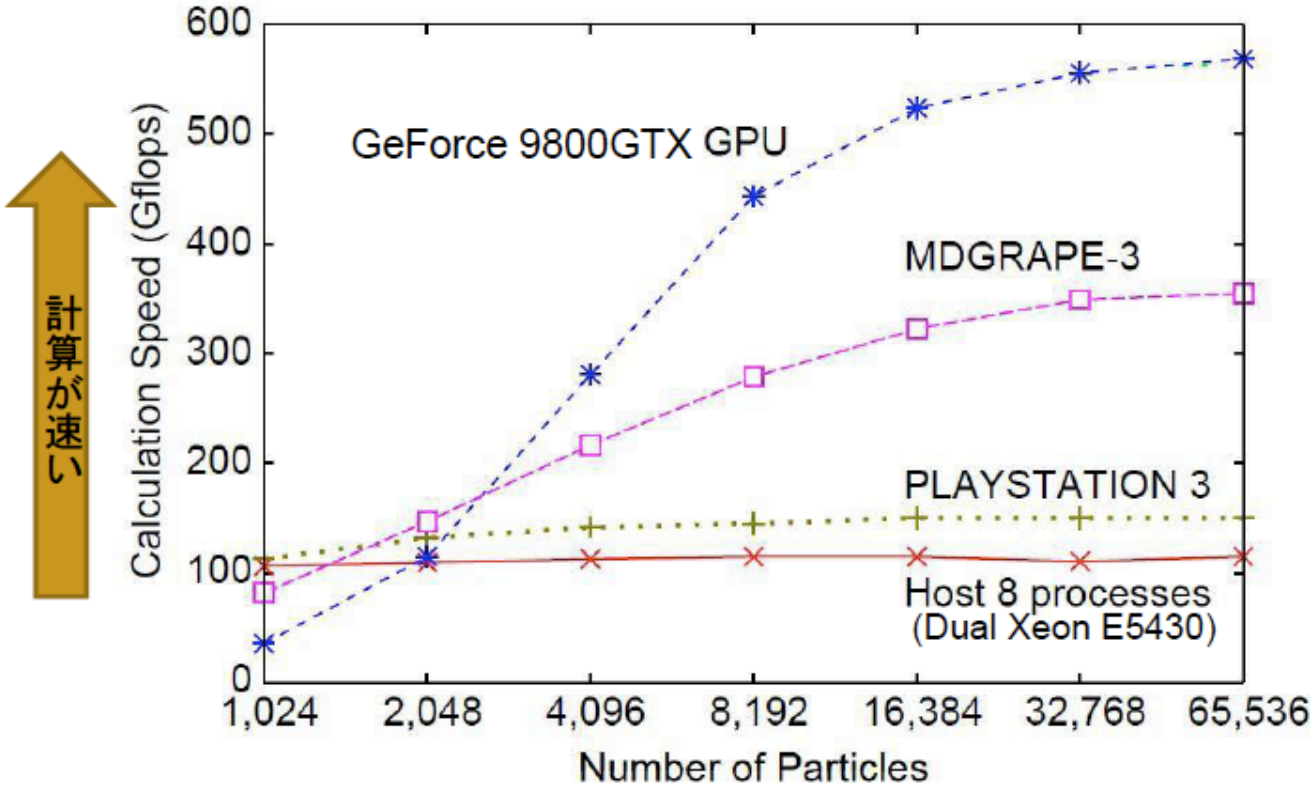

ClearSpeed X620


PLAYSTATION 3


GPU

重力多体問題での計算速度

- GPUとMDGRAPE-3は粒子数が少ない時に性能が落ちる



T. Narumi, et al., Proc. of PDCAT'08, pp. 143-150, New Zealand, (2008)
38 floating point interaction is assumed for pairwise gravity calculation

電通大（成見先生より）


DX X
GALAXY

ソフトウェアシンポジウム2009

アクセラレータの問題点

- 高い実行速度を得るためのプログラミングがかなり大変
  - １００～１０００程度の並列に対応する必要がある。
    - GPUの場合には、数千の並列が必要。
- 計算ユニットが不均一となる（Cell B.E)
  - PowerPC
  - SPE
- いろんな種類のメモリを使いこなす(GPU)
  - Global Memory, shared memory, texture memory, constant memory, local memory, register

若い人材に並列計算を身近に感じて貰う試みが必要

- Knoppix for MDGrapeで実施した環境を、GPU用に準備して、並列計算を簡便に試すことができる環境提供が必要→東工大（小西）、長崎大学（濱田）、電通大（成見）で提供開始。
- GT200系のGPUカードに対応した、Knoppix for CUDAのイメージを公開中。
- TESLA用 Knoppix イメージの配布準備中

理研から東工大に移ってみると

- TSUBAMEという非常に大規模で高速なスパコンが利用できる環境。
- 世界でもアクセラレータを積極的に取り入れたスパコン（ClearSpeed　SIMDアクセラレータ）
- 最近では、Teslaを導入して世界最初のGPGPUクラスターとなった。

GPUプログラミングができなければ、本来の性能を引き出せなくなってきた。→効果的な教育の必要性

国内のGPU関連研究者

氏名	組織	
青木尊之	東京工業大学	数値流体力学
秋山泰	東京工業大学	バイオインフォマティックス
遠藤敏夫	東京工業大学	HPCベンチマーク
小西史一	東京工業大学	バイオインフォマティックス
額田 彰	東京工業大学	数値計算（FFT)
松岡聡	東京工業大学	HPCアーキテクチャー
丸山 直也	東京工業大学	HPC
泰岡 顕治	慶応大学	分子動力学シミュレーション
成見哲	電気通信大学	分子動力学シミュレーション
濱田剛	長崎大学	自動並列化コンパイラ
萩原 兼一	大阪大学	画像処理
伊野 文彦	大阪大学	生体シミュレーション
滝沢寛之	滝沢寛之	プログラミング環境
森眞一郎	福井大学	可視化
泰地 真弘人	理化学研究所	HPC, 分子動力学シミュレーション
奥田洋司	東京大学	HPC

東工大はGPUコンピューティングでは大きく貢献

TSUBAME1.2の概要の紹介

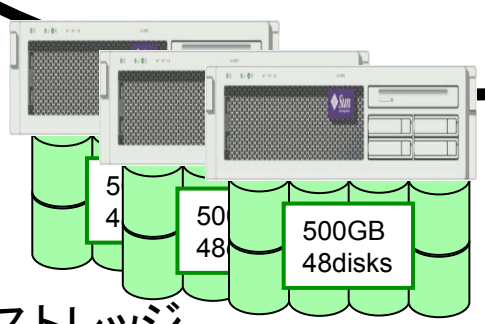
2006年4月東工大スパコン ”TSUBAME”
(スパコンとしての調達分)

Voltaire ISR9288 Infiniband x8
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)

10Gbps+外部
ネットワーク

2006年6月現在
アジア No.1, 世界No.7
38.18Teraflops
(Top500計測値)

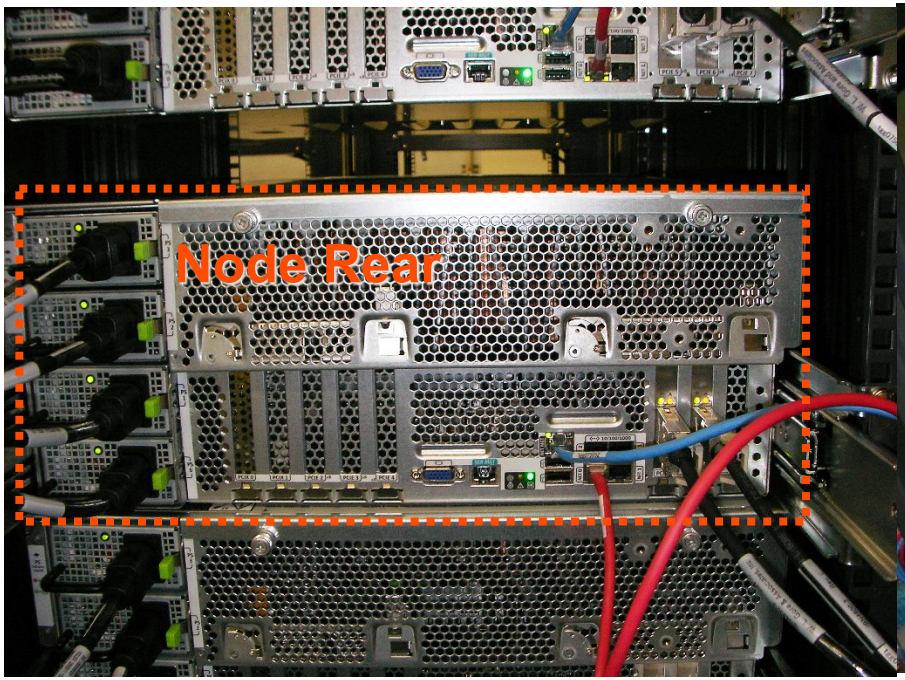
Sun/AMD高性能計算クラスタ
(Opteron Dual core 8-Way)
10480core/655ノード
50.4TeraFlops
OS(現状) Linux
(検討中) Solaris, Windows
NAREGIグリッドミドル



ストレッジ
1 Petabyte (Sun “Thumper”)
0.1Petabyte (NEC iStore)
Lustre ファイルシステム
>400Gbps

ClearSpeed CSX600
SIMD accelerator
360 boards,
30TeraFlops


Titech TSUBAME
~76+ racks
350m2 floor area
1.2 MW (peak)


Node Rear


Local Infiniband Switch
(288 ports)
Currently
2GB/s / node
Easily scalable to
8GB/s / node


~500 TB out of 1.1PB

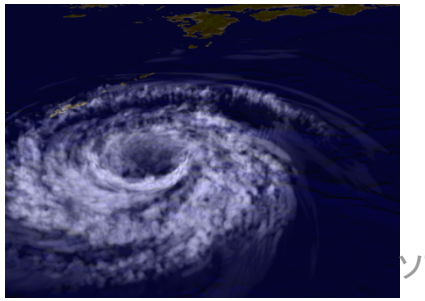

Cooling Towers (~32 units)

TSUBAMEの二年の運用成果：全目標達成

1. 東工大のシンボル：世界トップレベルの情報インフラ

＊Top500＊
4期連続日本一
4期連続性能向上
(世界初)

2. 研究推進：莫大な計算パワー・ストレッジ(1ペタバイト以上)・みんなのスパコン



3. 産学連携等の推進、大型プロジェクトへの呼び水、アライアンスを組む他大学計算ニーズホスティング

MITSUBISHI CHEMICAL
NUMERICAL TECHNOLOGIES
SUMITOMO CHEMICAL

文部科学省先端研究施設共用イノベーション創出事業【産業戦略利用】

Global COE
「計算世界観」

CompView

AMD
NEC
Sun microsystems
Microsoft

企業との包括Collaboration

Research Organization of Information and Systems
国立情報学研究所
National Institute of Informatics

NAREGI/NII-CSI
全国サイバーサイエンスインフラ
＊NAREGI 開発への貢献
＊阪大-東工大 NAREGI β2連携

4. 学内の分散した情報基盤の集約化・ホスティング

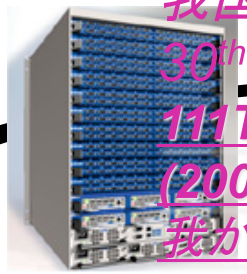
・TSUBAME「みんなのスパコン」
・新概念の課金利用法によるユーザ数増加 => 1300人へ倍増
・SE運用業務の追加(アプリ・性能評価・グリッド試験運用など)
・各種ITサービスのホスティング

TSUBAME 2008年4月、約一年前の姿
(東工大GSICセンターの複合システム)

Voltaire ISR9288 Infiniband x8
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)

10Gbps+External NW

Top500にて4期連続で
我国最速のスパコン
30th Top500@56.43TF
111TFlops Peak
(2008年6月67.7TF Top500
我が国で3位)



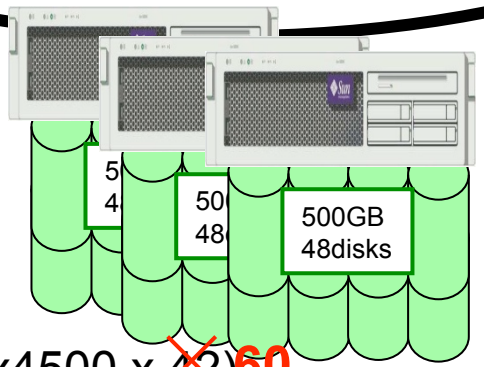
Sun x4600 (16 Opteron Cores)
~~32~64~~ GBytes/Node
32~128GB
(平成18年度末
NESTREシステム)
10480core/655Nodes
21.4TeraBytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

Unified Infiniband network

NEC SX-8i
(for porting)
(平成18年度ベクトル計算機追加)

"TSUBASA" 整数演算アクセラレータ
720 Cores, 8.1TF

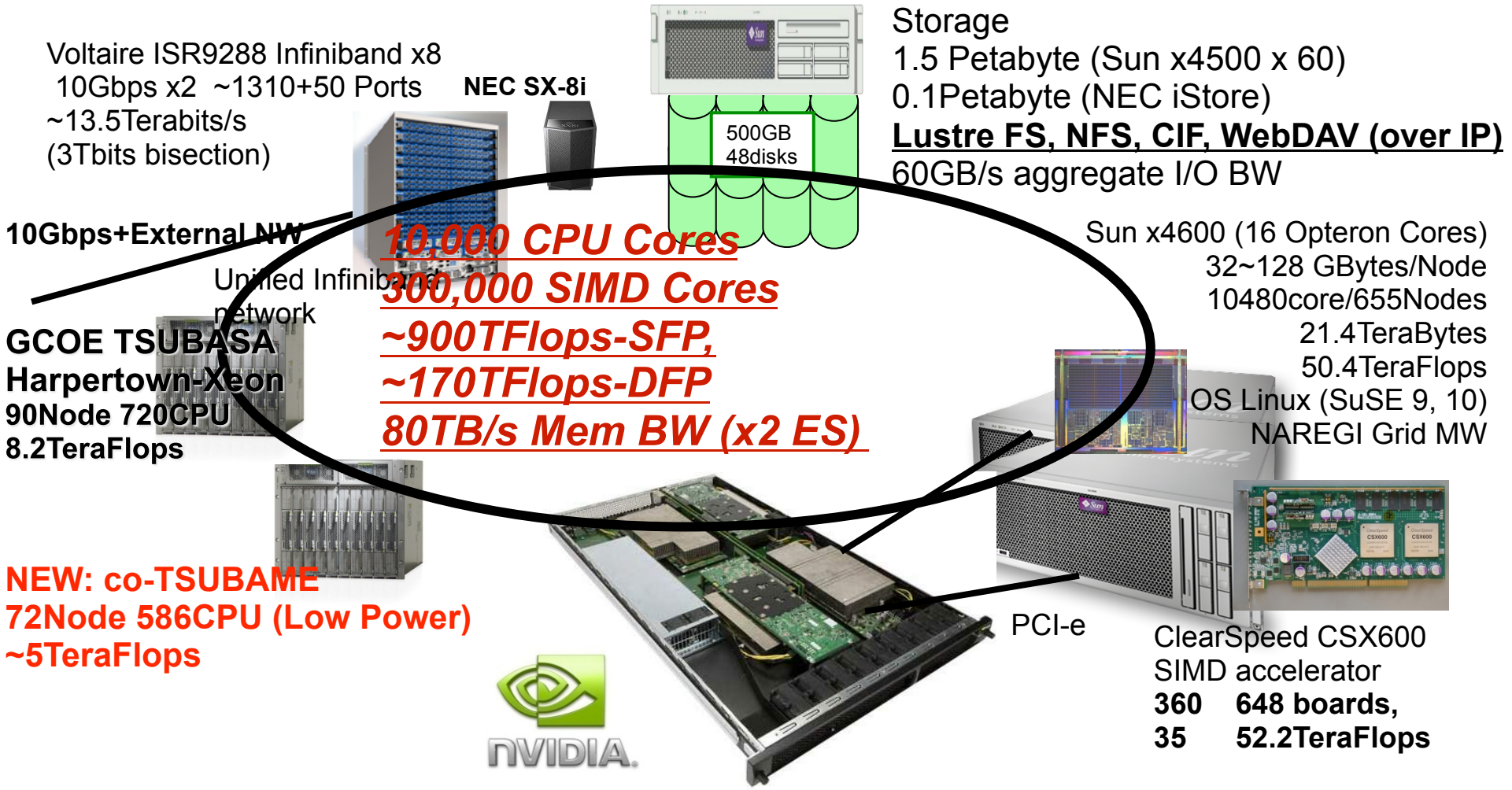
(平成19年度末
Global COE
「計算世界観」)



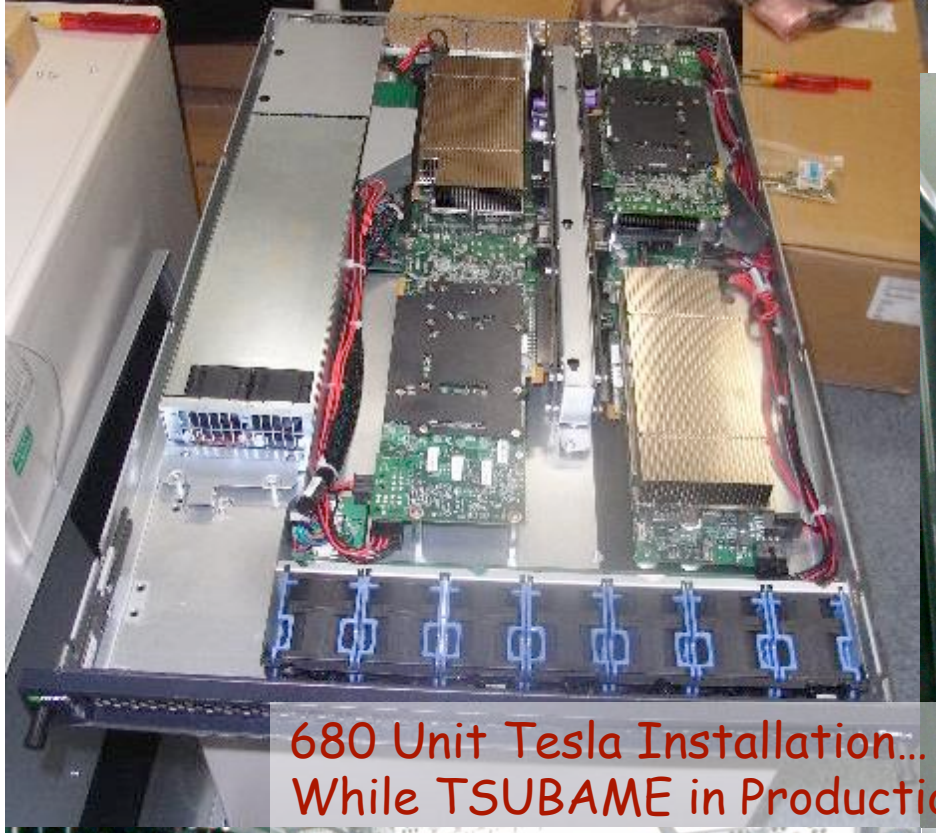
Storage
1.5PB
~~1.0 Petabyte~~ (Sun x4500 x ~~42~~)**60**
(平成18年度末
NESTREシステム)
0.1Petabyte (NEC iStore)
Lustre FS, NFS, CIF, WebDAV (over IP)
60GB/s ~~50~~GB/s aggregate I/O BW

ClearSpeed CSX600
SIMD accelerator
~~360~~ **648 boards,**
~~35~~ **52.2TeraFlops**
(平成19年度
分子動力学
アクセラレータ)

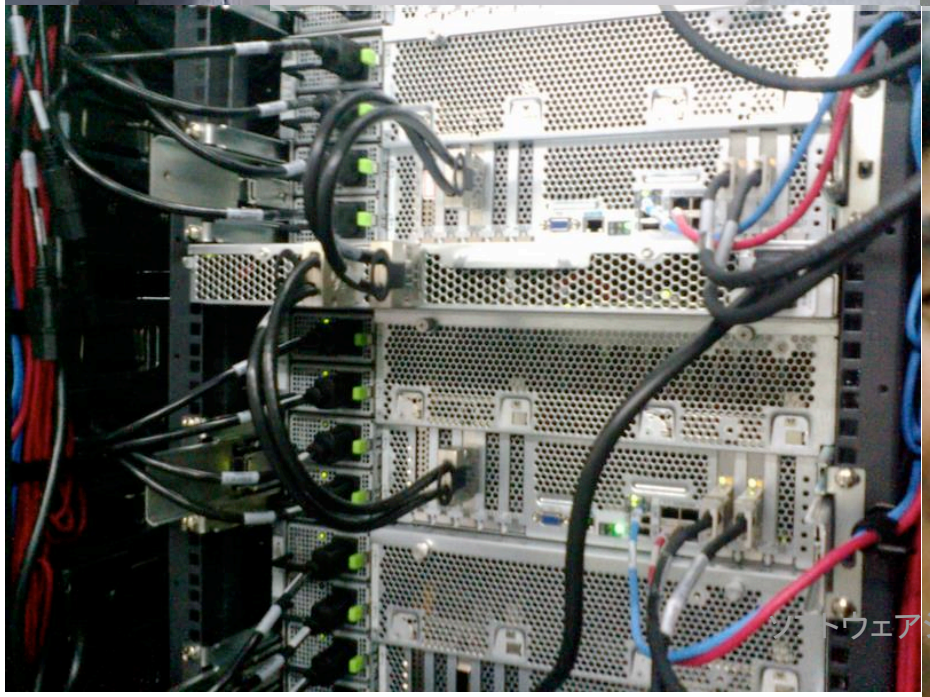
TSUBAME 1.2への進化=>GPUの試験的追加



Nvidia Tesla S1070: 170台, 総計 680カード
High Performance in Many BW-Intensive Apps
10% power increase over TSUBAME 1.0 (130TF SFP / 80TF DFP)



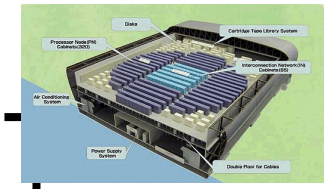
680 Unit Tesla Installation…
While TSUBAME in Production Service (!)



- 最新の2月号
- GPU関係の記事4篇
  => 4名の著者
  - 松岡　聡
  - 遠藤　敏夫
  - 青木　孝之
  - 小西　史一（共著）

TSUBAME2.0へ向けた性能向上

地球シミュレータ ⇒ TSUBAME 4年間
40倍のダウンサイズ



TSUBAME ⇒
TSUBAME 2.0
4年間で40倍???

TSUBAME2.0
??? PF (2010-1H)

Japanese NLP
>10PF(2012)
US >10P
(2011~12)

US/EU
Petascales
(Peak)
(2008)

US NSF/DoE
(2010)

U-Tokyo, Kyoto-U,
Tsukuba T2K 61-140TF (2008)

Others

BlueGene/L
360TF(2005)

TSUBAME 1.2
Upgrade
(3Q2008) 170TF

TSUBAME 1.1 Upgrade 111TF, 1.6
PB, 128GB nodes (1Q2008)

TSUBAME 1.0 80TF (1Q2006)

Earth Simulator
40TF (2002)

Titech
Campus Grid
1.3TF

KEK 59TF
BG/L+SR11100

10PF
1PF
100TF
10TF
1TF

2002
2004
2006
2008
2010
2012

並列計算における教育

- Knoppix for HPC/HTCで作成してきたインスタント環境は、TSUBAMEの環境の縮図として教育利用が可能。

- GPU教育プログラムでの利用を準備中
  - 情報処理学会主催のGPUチャレンジでのサンプルコードや、入賞コードの共有メディアとして活用

GPU教育プログラム

- 遠藤敏夫・小西史一・松岡聡
- 計算数理実践-HPC-(Advanced application of Computing and Mathematical Sciences-HPC-
- 高性能計算(HPC)に関する実践的な知識，技術を提供することを目的とする．座学だけではなく，TSUBAMEスパコン・CompView TSUBASAクラスタを用いたHPCプログラミングの実習を行う．MPI, OpenMPなどの標準的な並列プログラミング環境に加え，近年注目を集めているGPU上のプログラミング環境であるCUDAについても講義・実習を行う.

授業スケジュール

内容	
1. Introduction to Pragmatic HPC	
2. C Language	C言語のおさらい
3. OpenMP Part 1	
4. OpenMP Part2	
5. MPI Part 1	メッセージパッシングの基礎 SPMDプログラミング
6. MPI Part 2	ノンブロッキング通信
7. MPI Part 3	MPI-2のリモートメモリアクセス 数値計算プログラムを用いた実習
8. CUDA Part 1	GPUアーキテクチャ CUDA言語入門
9. CUDA Part 2	CUDAにおけるスレッド・メモリ階層
10. CUDA Part 3	行列積・N体問題を用いたプログラム例
11. CUDA Part 4	CUDAプログラムの高速化のための知識
12. Application Part 1	Bioinformatics 紹介
13. Application Part 2	Dynamic Programing
14. Application Part 3	Smith-waterman

OSS分野とHPC分野との接点

- 現在、TSUBAMEにとって必要なのは品質の良いGPUコード

- 特に、数値演算ライブラリなどが重要
  - 利用者が、ランタイムにロードして利用できるのがベスト（実行環境に応じて、切り替えられ、最適化が行われるもの）
  - 東工大では、高速な3D-FFTライブラリの開発