SHEEPDOGクラスタ運用と広域化への試み

島 慶一 <shima@wide.ad.jp>

第4回クラウドストレージ研究会

2011年12月8日

WIDE PROJECT

IIJ INNOVATION INSTITUTE

動機

- WIDE Cloudの活動

- 地理的/組織的に分散配置されたHVを相互運用するための運用実験基盤

- ハイパーバイザーはubuntu/KVMを利用

WIDE PROJECT

IIJ INNOVATION INSTITUTE

動機

- 広域運用の課題はネットワークとストレージ
  - 広域VLAN、L2TP、NEMO BSなどを使ったネットワーク仮想化 (今後、vxlan、LISP、openflow?などを検討)
  - 広域NFS、広域iSCSI (今後、sheepdog、nbdなどを検討)

WIDE PROJECT

IIJ INNOVATION INSTITUTE

動機

- 法定点検のため組織単位で停電とか
- でもサービスは止めたくない
- 突然の障害にも対応できたらいいな
- ついでにDRとかにも使えないかな。。。

動機

- 法定点検のため組織単位で停電とか
- でもサービスは止めたくない
- 突然の障害にも対応できたらいいな
- ついでにDRとかにも使えないかな。。。

みなさんご存知の

- Sheepdog
- 森田氏 (NTTサイバースペース研究所)
- 対称型の分散ストレージ
- 仮想化環境での利用を念頭に置かれた設計

知りたい事

- 拡張性
- 信頼性
- 保守性
- 性能

STARBED

- 大規模ネットワーク実験環境 (運用 NICT)
- 仮想じゃないクラウドのようなもの
- 960台のサーバー群とそれらを接続するスイッチ
- http://www.starbed.org/

WIDE PROJECT

IIJ INNOVATION INSTITUTE

現時点での目標

- SheepdogクラスタをWIDE Cloudのストレージ基盤として導入
- 広域に分散されたSheepdogクラスタの運用
- iSCSIインターフェースによる仮想マシンへのストレージ提供

練習

- やってみないことには感触がつかめないので
- StarBEDのノードを50台ほど借用して動かしてみる

構成 (ノード)

- Cisco UCS 200 M2
  - Intel Xeon X5670 (6 コア) x 2
  - 48GBメモリ
  - SATA 500GB x 2
  - 6 NIC (Intel Gigabit double、 Broadcom 5709 quad)

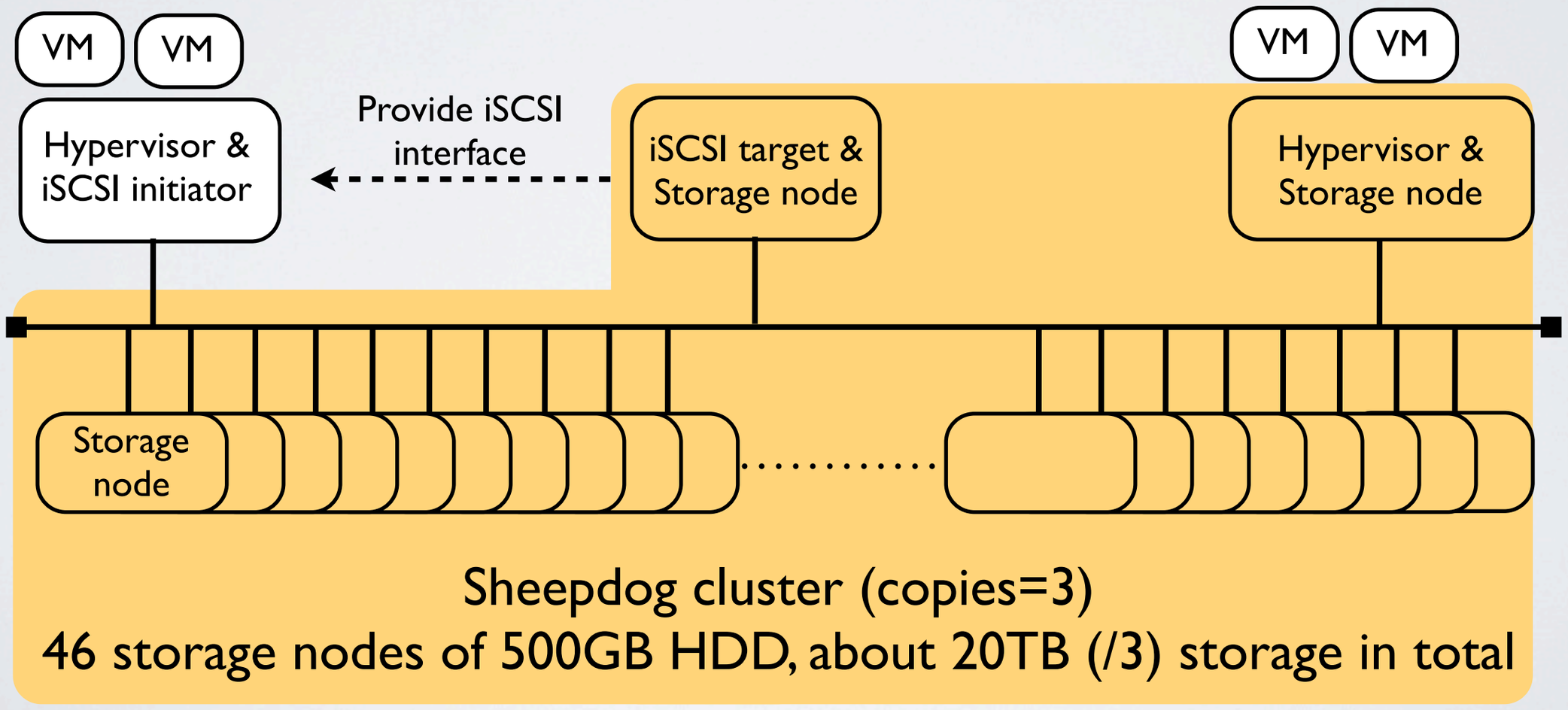
WIDE PROJECT

IIJ INNOVATION INSTITUTE

構成 (ネットワーク)

Virtual machines running outside of the sheepdog cluster using iSCSI interface

Virtual machines running directly on the sheepdog cluster



できたこと

- 46台でのクラスタ構成と運用
- 既存クラスタへのノードの順次追加
- クラスタ内での仮想ディスク構成
- クラスタ内での仮想マシン稼働

WIDE PROJECT

IIJ INNOVATION INSTITUTE

はまったこと

- そもそものStarBEDの使い方 :-)
- master branchの不安定さ
  - devel branchに変える事である程度解消
  - 0.3リリースに期待
- 壊れたクラスタの修復ができないこと
  - なにか起きるたびに0からやりなおし

WIDE PROJECT

IIJ INNOVATION INSTITUTE

不安な点

- Sheepdogクラスタからストレージノードの切り離し
  - クラスタが安定するタイミングが分からない
- 稀に各ストレージノードが把握しているグループメンバーの一貫性が壊れる
- iSCSIターゲット機能が不安定

残念ながら

- StarBEDではLAN内での運用しかテストできなかった
- 時間的な制約のため (StarBEDの制限ではなく)
- もっと大規模、かつdelayを挟んだ広域 (エミュレーション) での実験、評価は今年度内に予定

WIDE PROJECT

IIJ INNOVATION INSTITUTE

広域実験

- WIDE CloudとJGNの機材を用いた実環境でのテスト

- 合計13台のSheepdogノード

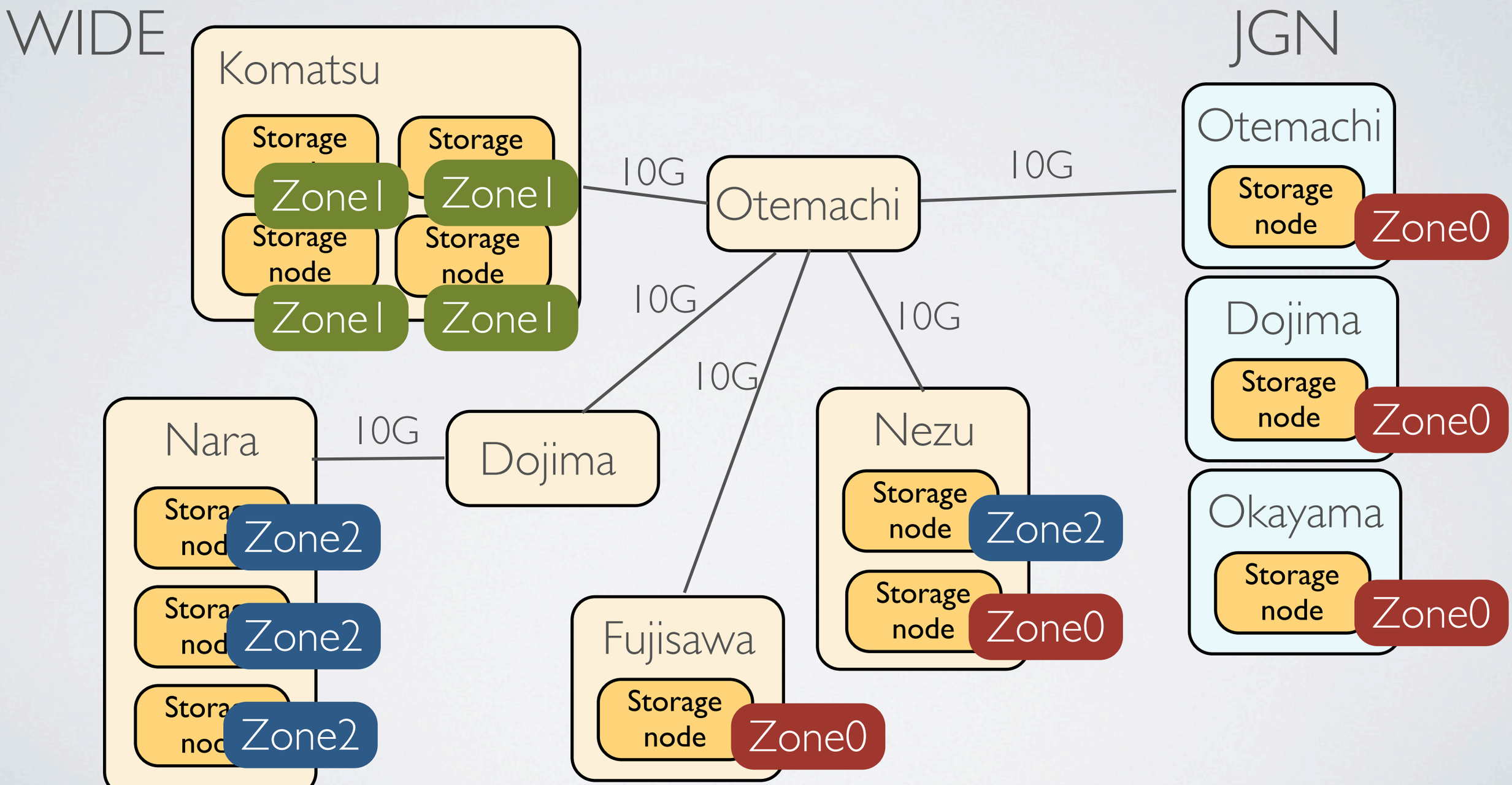
- 全国に配置されたストレージノード

- Sheepdogのzone機能による地理的集約

WIDE PROJECT

IIJ INNOVATION INSTITUTE

WIDE CLOUD & JGN



BONNIE++での仮計測

- ハイパーバイザー
  - Intel Xeon L5520 @ 2.27GHz (4コア x 16), 48GBメモリ, HP NC522SFP Dual Port 10GbE
  - ubuntu 10.04.3 LTS
  - Sheepdog version 0.2.4_125_gc74e340
- 仮想マシン
  - QEMU version 0.13.0
  - ubuntu 10.04 LTS
  - 1 VCPU, 1GBメモリ, 16GB sheepdogディスク

WIDE PROJECT

IIJ INNOVATION INSTITUTE

BONNIE++での仮計測

	Result	Latency
Seq Write (Char)	638K/sec	16823us
Seq Write (Block)	4344K/sec	24527000us
Seq Rewrite	2227K/sec	15843000us
Seq Read (Char)	2267K/sec	149000us
Seq Read (Block)	11453K/sec	74419us
Random Seek	101.5/sec	9631000us
Seq Create	4654/sec	766us
Seq Delete	5514/sec	890us
Random Create	4762/sec	650us
Random Delete	7724/sec	219us

WIDE PROJECT

IIJ INNOVATION INSTITUTE

まとめ

- Sheepdogはおもしろい
- だが、今、運用できるかと問われると、まだ難しそう
  - 障害時の対処が困難
  - HVとストレージノードのカップリング
  - ノードの追加削除に伴うシステムの動きの定量化
  - クラスタのヘルスチェック機能
  - 大規模化による制御パケットの氾濫の心配

WIDE PROJECT

IIJ INNOVATION INSTITUTE