

大規模サーバ運用者の苦悩 (笑)

CHAOS編



SRSさくらインターネット
代表取締役 田中邦裕

大規模サーバとは(1)

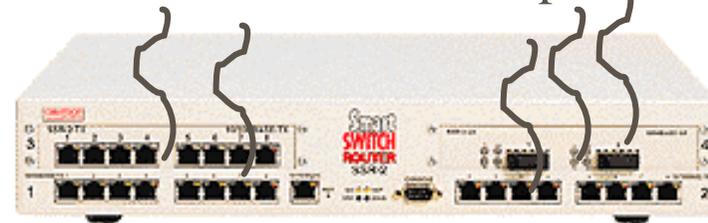


想像できないことが生じるサーバ群

夜間にサーバが能力を発揮しすぎて、60%だった電源使用率がフルに
当然ブレーカが……



1Mbps位しかなかったトラフィックが1週間後に100Mbpsになったとか
当然ルータが……



16000ファイル開けるはずのサーバなのにFile open できないとか……

130,000あるmbufがいっぱいとか……

あれとか……

これとか……

大規模サーバとは(2)



エンドユーザとの心理戦をしなければならないサーバ群

ジャイアント馬場が亡くなったときにLoadAverageが100を超えたサーバ
ファンサイト掲示板への書き込みが暴発

ワールドカップの際に10倍のトラフィックとなった某匿名掲示板サーバ
代表チームと一緒にサーバと戦って…。

停止したら大変と言われながら良く止まるサーバ群！？

1日 1億ヒットのシステムを作るための予算は1000万ヒットのシステムの
10倍かけても足りないのは……なぜ？

大規模サーバ 2種類の定義

自社コンテンツを提供する場合

利用者は常識的にサーバを利用する

利用者はサーバ管理者に遠慮する？

ファイアウォールで守る

LAW系

← Yahoo!

PlayOnline

民田さんところ

他社(者)コンテンツを預かる場合

利用者は非常識にサーバを利用する

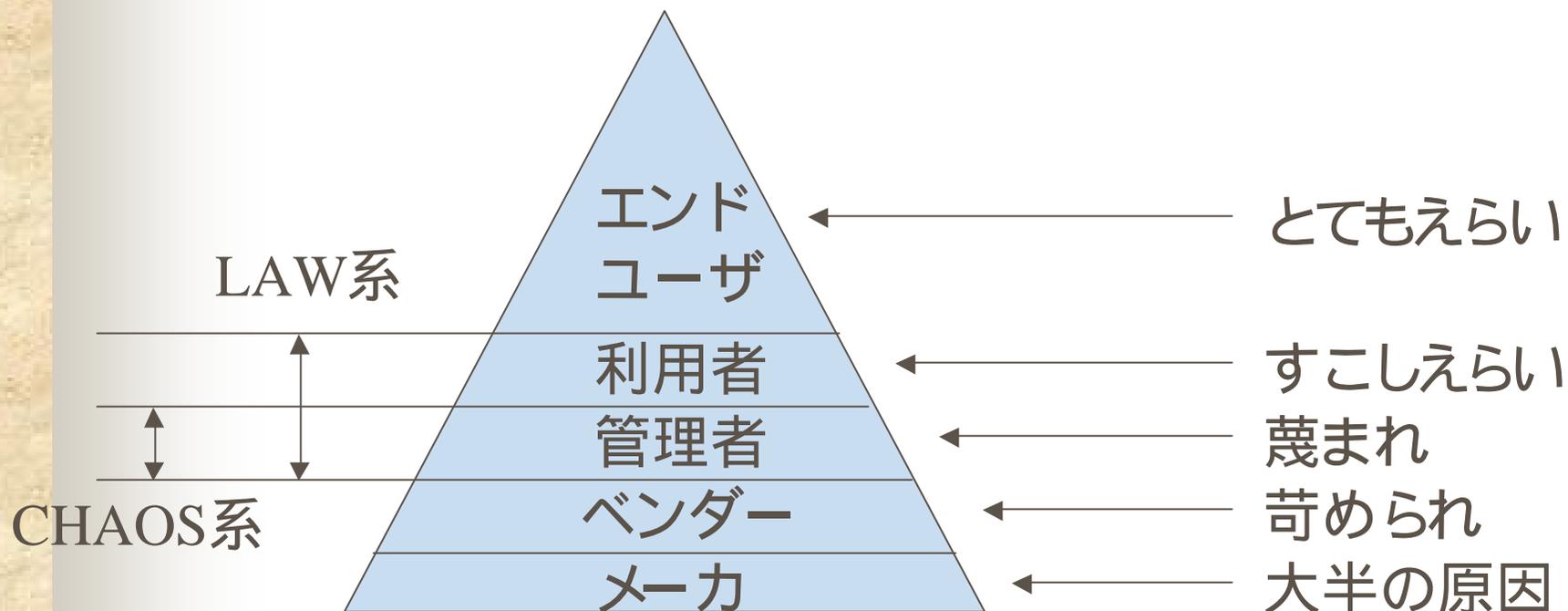
利用者はサーバの限界を試したがる

ひたすらパッチをあて続ける

CHAOS系

うち →

サーバ 5階層モデル



では、CHAOS系大規模サーバの苦悩をつづってゆきます

サーバの概況



ウェブホスティング用サーバ

FreeBSD 4.5-RELEASE-p19300台

サービス用サーバ (キャッシュ・DNS・メール中継等)

FreeBSD 4.x40台

専用サーバ

FreeBSD 3.x350台(概数)

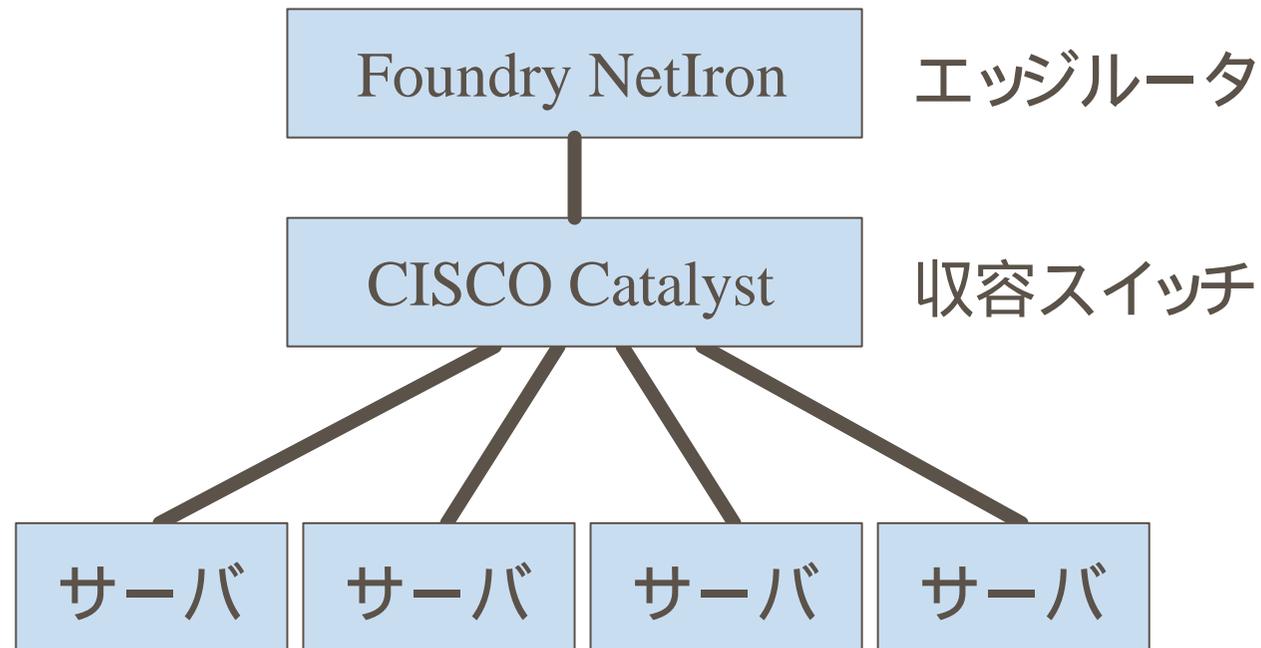
FreeBSD 4.x200台(概数)

RedHat Linux800台(概数)

その他、たくさんのハウジング・コロケーション・・・

ウェブホスティング用サーバ

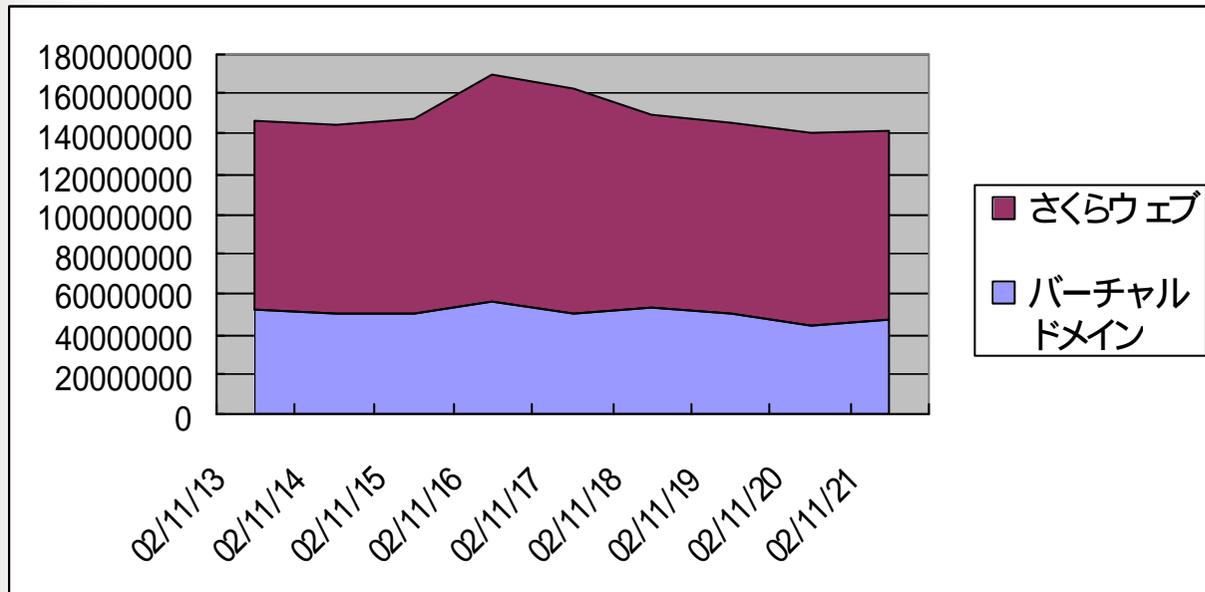
アカウントを登録し、そのアカウント毎にquotaをかけて、ウェブスペースをレンタルする



アクセスの概況

WWWサーバのヒット数でいうなら

ウェブホスティングサービス向けのアクセスは1日1億5,000万ヒット程度

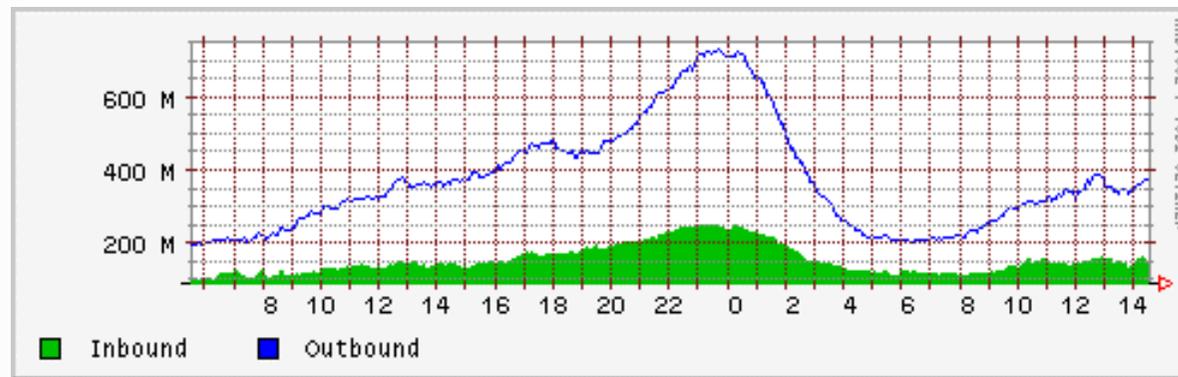


	バーチャルドメイン	さくらウェブ
11月21日	47315421	94006126
11月20日	44818217	96159073
11月19日	49989802	95321642
11月18日	52475018	96860238
11月17日	50774562	111568576
11月16日	57042924	112526180
11月15日	50425787	97069520
11月14日	50333528	93827539
11月13日	51750152	95147016

アクセスの概況

レートで言うなら

ウェブホスティング及び専用サーバで、夜間に1Gbps程度



上記のグラフは東京第1及び第2ネットワークセンターの合計
(主にウェブホスティングサーバとサーバハウジング)

専用サーバ

サーバをそのままレンタルするサービス
オプションでセキュリティアップデート



ウェブホスティング編

こんなことよくあるよね

自分のホームディレクトリ内に/usr以下を構築

ホスティングサーバなのに、muleやgccが入っているのは当然のこと
まあ約款違反ではないので黙認。
たまにファイル数の限界でQuotaに引っかかっている人がいる模様

Apacheが入ってて8080番ポートで動作させたり、SMTP動かしたり
最近やっとファイアウォール入れました。
セキュリティのためでなく、常駐プロセスを削減させるために

CPUをとりあえず使ってみる

パスワード解読は自宅でやってください・・・。

サーバでgccを再構築したりするのはかんべん

CronでNAMAZUのIndexを作り直した上、Digest配送したりなんか・・・

まるで、専用サーバのように・・・

ウェブホスティング編

よくあって困る状態とは

ゾンビプロセスをいっぱい作成する

他のサーバでは問題なかったのに」が常套句

いっぱい fork した挙句、プロセス暴走

おなじく 他のサーバでは問題なかったのに」が常套句

ファイルロックのやり方がしょぼく、プロセス残留

やっぱり 他のサーバでは問題なかったのに」が常套句

テンポラリファイルを削除せず、i-Node を使いまくる

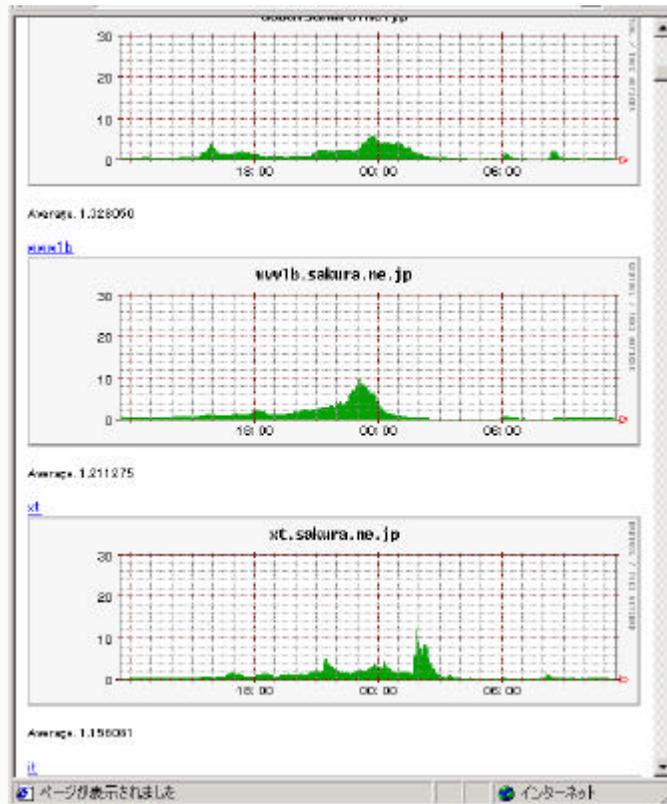
なにがあっても 他のサーバでは問題なかったのに」が常套句

こんなことを起こす可能性のあるサーバが、何百台も・・・

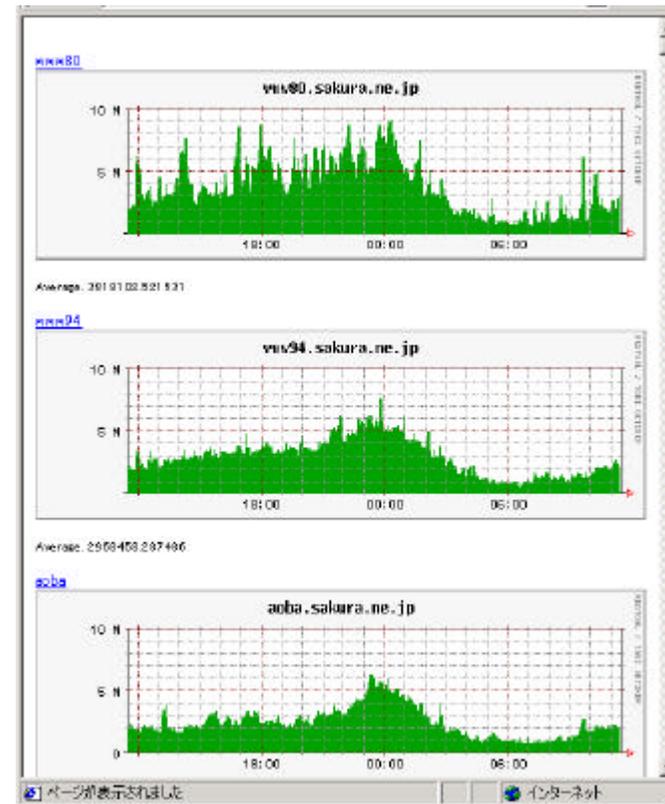
ウェブホスティング編 傾向と対策(1)

2つのグラフの具合で、サーバ職人が負荷状況を推測する。

ロードアベレージを観測
ランキング化し統計を取る

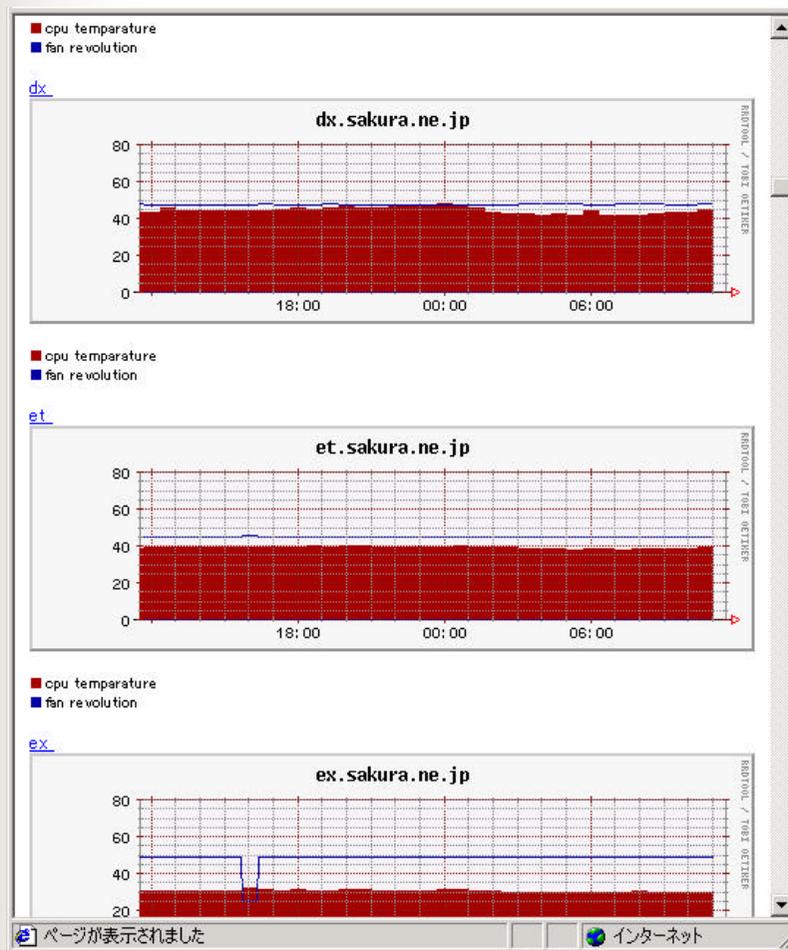


トラフィックも同様に観測
ランキング化し統計を取る



ウェブホスティング編 傾向と対策(2)

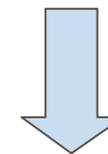
2つのグラフの具合で、サーバ職人がサーバ寿命を推測する。



CPUの温度とファン回転数をランキング化し統計を取る。
ファンが故障するとグラフが面白めだったけど、全て山洋にすると激減

失敗談も・・・

職人ならCPU温度で暴走したプロセス数を推測できるのではないかと教え



あえなく失敗。
仕方が無いので、ロードアベレージと次ページの方法で推測することに

ウェブホスティング編 傾向と対策(3)

新米管理者がレポートを見て、サーバの危機を職人に報告

実行時間の長いプロセスを
一定時間毎に報告

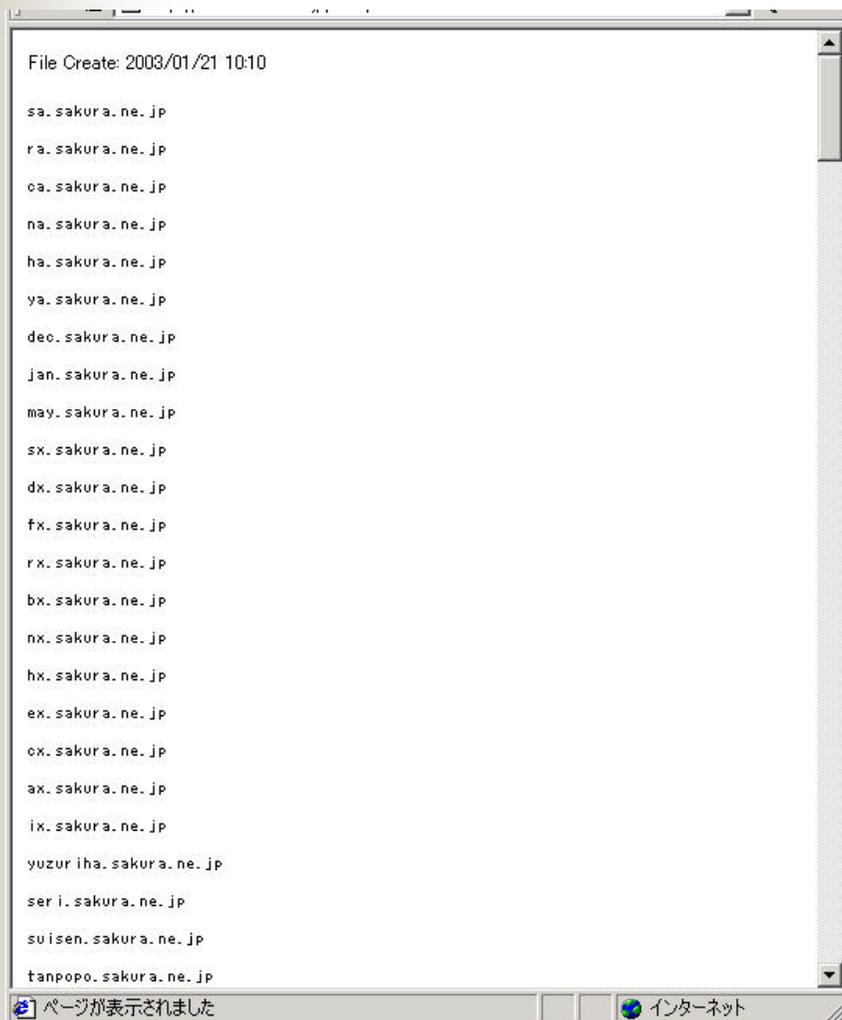
ディスク容量を報告
80%を超えると赤い表示

```
File Create: 2003/01/21 09:09
sa.sakura.ne.jp
ra.sakura.ne.jp
  1211 0.0 1.6 4632 4108 p0-5 3Jan03 3:47.36 nedaka_C1Pauto/
ca.sakura.ne.jp
  72286 11.0 0.0 0 0 71 Z 9:50AM 0:00.00 (par 1)
  72181 0.0 0.7 2424 1804 ?? R 9:50AM 0:00.15 /usr/local/bin/pc
sa.sakura.ne.jp
ba.sakura.ne.jp
sa.sakura.ne.jp
fad.sakura.ne.jp
jan.sakura.ne.jp
  29863 0.0 0.4 1420 960 71 1c 9:28AM 0:08.73 httpd: p0026-1p0t5x0
way.sakura.ne.jp
sk.sakura.ne.jp
  60111 10.0 0.9 2762 2162 17 8 9:50AM 0:10.89 /usr/local/bin/
  6070 0.0 1.9 6584 4556 p0-1 9Jan03 6:15.40 /usr/bin/parl -
da.sakura.ne.jp
  8578 0.0 0.7 2420 1860 71 5 9:58AM 0:08.04 /usr/local/bin/parl
fa.sakura.ne.jp
  78398 0.0 0.9 2772 2208 p0 1c 7:53AM 0:08.24 /home/ai11/bin/zh-
  78568 0.0 0.2 886 892 p0 1+ 7:54AM 0:08.14 1211 -1 8121loop_no
rx.sakura.ne.jp
kx.sakura.ne.jp
sk.sakura.ne.jp
  54289 0.0 0.5 1880 1200 17 8 9:47AM 0:10.82 parl kanzi.pl
kx.sakura.ne.jp
  75784 3.0 0.3 1224 890 17 8x 9:50AM 0:10.29 popper -xR
  75766 0.0 1.5 4506 2972 17 R 9:50AM 0:10.17 /usr/local/bin/par
```

www23.dns.ne.jp	12%	70%	54%
www24.dns.ne.jp	12%	66%	27%
www25.dns.ne.jp	12%	64%	20%
www26.dns.ne.jp	12%	79%	18%
www27.dns.ne.jp	12%	54%	27%
www28.dns.ne.jp	12%	66%	44%
www29.dns.ne.jp	12%	25%	36%
www30.dns.ne.jp	20%	80%	23%
www31.dns.ne.jp	12%	82%	40%
www32.dns.ne.jp	11%	45%	32%
www33.dns.ne.jp	18%	95%	82%
www34.dns.ne.jp	12%	57%	82%
www35.dns.ne.jp	12%	80%	80%
www36.dns.ne.jp	12%	27%	45%
www37.dns.ne.jp	12%	45%	30%
www38.dns.ne.jp	12%	60%	60%
www39.dns.ne.jp	12%	21%	40%
www40.dns.ne.jp	12%	19%	20%
www41.dns.ne.jp	11%	19%	44%
www42.dns.ne.jp	11%	15%	20%
www43.dns.ne.jp	12%	25%	19%
www44.dns.ne.jp	11%	22%	58%
www45.dns.ne.jp	11%	40%	88%
www46.dns.ne.jp	12%	62%	24%
www48.dns.ne.jp	12%	82%	26%
www49.dns.ne.jp	22%	44%	62%
www50.dns.ne.jp	12%	90%	36%
www51.dns.ne.jp	14%	83%	30%
www52.dns.ne.jp	12%	32%	59%
www53.dns.ne.jp	12%	22%	17%
www54.dns.ne.jp	12%	25%	25%
www55.dns.ne.jp	11%	26%	10%
www56.dns.ne.jp	11%	20%	13%
www57.dns.ne.jp	12%	54%	60%
www58.dns.ne.jp	11%	20%	27%
www59.dns.ne.jp	11%	31%	65%

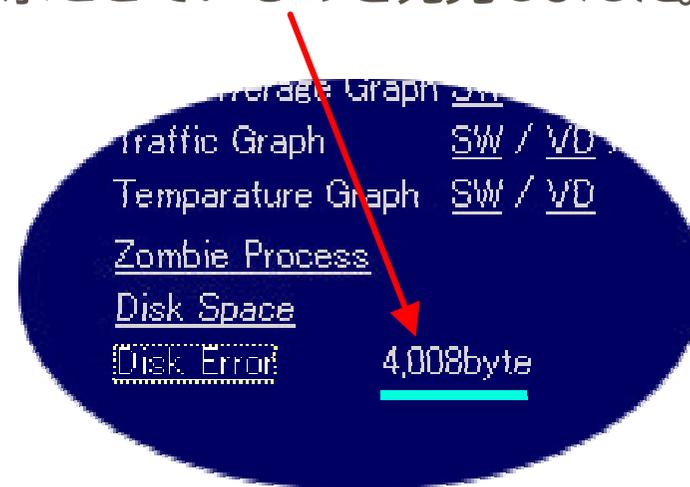
ウェブホスティング編 傾向と対策(4)

新米管理者がレポートを見て、サーバの危機を職人に報告



残念ながら?、エラーを出しているサーバは無かったので左の図は参考出品です

ところで、サーバ職人は関西人なので「面倒くさがり」で「いらち」です。なのでIndexページでレポートのバイト数を表示させているのを発見しました。



専用サーバ編

こんなことよくあるよね

VPNのサーバを構築

SINETの商用・海外向けが遅いので、専用サーバを経由させ。
IPアドレスを64個要求され、後日用途を調べていると、専用サーバより
先までtracerouteが通る。あれっ、専用サーバなのにルーティングしてる？

IPアドレスを多く要求する人の大半が、VPN用途
データベースのバーチャルホストが一般的になってきたため、
IPアドレスを要求する人の大半が、サーバ外で利用しているようだ。

...

....

まるで、コロケーションのように・・・

専用サーバ編

よくあって困る状態とは

最近、サーバが重いのですが……

サーバがクラックされているようですけれども…。

キャッチホンがかかってきて……

こういうのは、hosts.allowを書いている時が多い。
とりあえずサポートにALL:ALL:allowと変更してくれとの依頼が。

IP追加しようとしたらサーバが止まったぞ……

ルータ君からIP重複させられたぞと連絡が届くと共に、監視サーバからデフォルトルータのIPが変わったと通知が来る。慌てて対象ポートをシャットダウン。「defaultなんとかにIPを入力したのになぜダメなんだ」と



予知でき防止できない状態

ところで、対策をしたくても、完璧な対策ができないことも・・・

強力なウイルスでたらしいよ

いつもよりディスクの空き容量を増やしておこう

3日後 :Quota Exceededがいっぱい。まあFile System Fullよりましかな

重大なセキュリティホールでたらしいよ

パッチを作って専用サーバユーザに連絡するのだ

サポートへのメール :tmp/にあるドットをつくプログラムは何でしょうか？

インターネットの縮図がここにありますね・・・

ウェブホスティング用サーバ サーバ設定の内容

```
defaultrouter="210.188.226.1"  
hostname="?.sakura.ne.jp"  
ifconfig_fxp0="inet 210.188.226.? netmask 0xfffff80 media 10baseT/UTP mediaopt full-duplex"  
nfs_reserved_port_only="YES"  
sendmail_enable="YES"  
sshd_enable="YES"  
syslogd_flags="-ss"  
firewall_enable="YES"  
firewall_script="/etc/ipfw.cfg"  
tcp_restrict_rst="YES"  
tcp_drop_synfin="YES"  
portmap_enable="NO"  
ntpdate_flags="dns5.sakura.ad.jp"  
ntpdate_enable="YES"  
enable_quotas="YES"  
check_quotas="YES"
```

```
•sshd  
•inetd  
  •qpopper  
  •ftpd  
  •telnetd  
•sendmail  
•cannaserver  
•apache
```

ウェブホスティング用サーバ

アカウントモデル

tanaka	グループ users
nakamura	
yamada	
taro	
⋮	

ウェブサーバ

ユーザ www

グループ www



ディレクトリのパーミッション

/var/log - root/wheel - 750

/home/??*/ - ??*/users - 705

/usr/local/apache/ - root/www - 750

ウェブホスティング用サーバ

バックアップ

バックアップは行わないと銘打っているものの、取らないわけには行かない。

120MB(サービス平均) × 100アカウント(平均) × 250台 = 約3TB

もちろん、全員が全容量を利用するわけではないが・・・。

現在500GBのストレージを3台用意

毎日朝の3:41と5:41に、日替わりでいろいろなサーバをバックアップ
バックアップにはrsyncを利用

概ね1～2週間に一度フルバックアップされる上位サービスは5日毎

rsyncのコマンドラインオプション

```
rsync -rlptgovz --delete --exclude '.mail' --exclude '.log' -e ssh server:/home /backup/7/server
```

ウェブホスティング用サーバ

バックアップ

ストレージは、80GB (ファイルシステムにすると70GBちょっと)のEIDEのHDDを7台搭載して実現。

システム用1台(4GB)と合わせて8台のIDEを利用
システム用1台とバックアップストレージのうち3台は、オンボードのインタフェース
残りの4台はPromise ATA66
を利用した

今なら120GBのHDDも
出まわっているので、
併せて実験中

#	df	Filesystem	1K-blocks	Used	Avail	Capacity	Mounted on
		/dev/ad0s1a	127023	42014	74848	36%	/
		/dev/ad0s1f	2808958	967864	1616378	37%	/usr
		/dev/ad0s1e	635183	57907	526462	10%	/var
		procfs	4	4	0	100%	/proc
		/dev/ad1s1c	77580665	62583398	8790814	88%	/backup/1
		/dev/ad2s1c	77580665	58210893	13163319	82%	/backup/2
		/dev/ad3s1c	77580665	60655425	10718787	85%	/backup/3
		/dev/ad4s1c	77580665	48373799	23000413	68%	/backup/4
		/dev/ad5s1c	77580665	39426245	31947967	55%	/backup/5
		/dev/ad6s1c	77580665	5522277	65851935	8%	/backup/6
		/dev/ad7s1c	76607880	26561192	43918058	38%	/backup/7

セキュリティ対策

セキュリティホールはOSには付き物といえるが、当社の場合お客様が直接触ることの出来るサーバである為、対処には敏感にならざるを得ない。

Updateの方法

たいていの事例では、sshを利用して修正バイナリを送りこむ
(複数のOSバージョンにまたがる場合にはバイナリを複数用意する)

Updateのスケジュール

外部からの攻撃が可能なものに付いては発表から概ね24時間以内に全サーバを対処。内部からの攻撃が可能なものについても極力同様に対処する。

以前は、外部からの攻撃が可能なもの以外は72時間以内の対処としていたものの、とあるサーバのログインユーザが「お手軽クラックツール」を利用して実験を行うという事例が報告され、対応を強化

バージョンアップ

お客様がご利用中のサーバは手軽にアップデートできない

現行の方法

新しいバージョンのFreeBSDをインストールしたサーバを用意しておき、古いほうのサーバよりコピーする。
移動時には、1時間程度のメンテナンス時間を頂き設定を移し替える。

検討中の方法

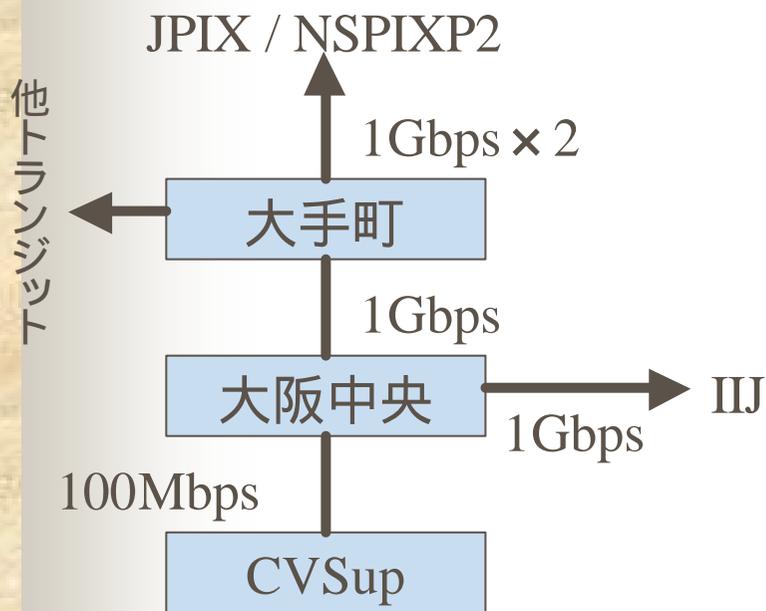
CVSupを利用する

常に最新のバージョンにできるメリットがある。

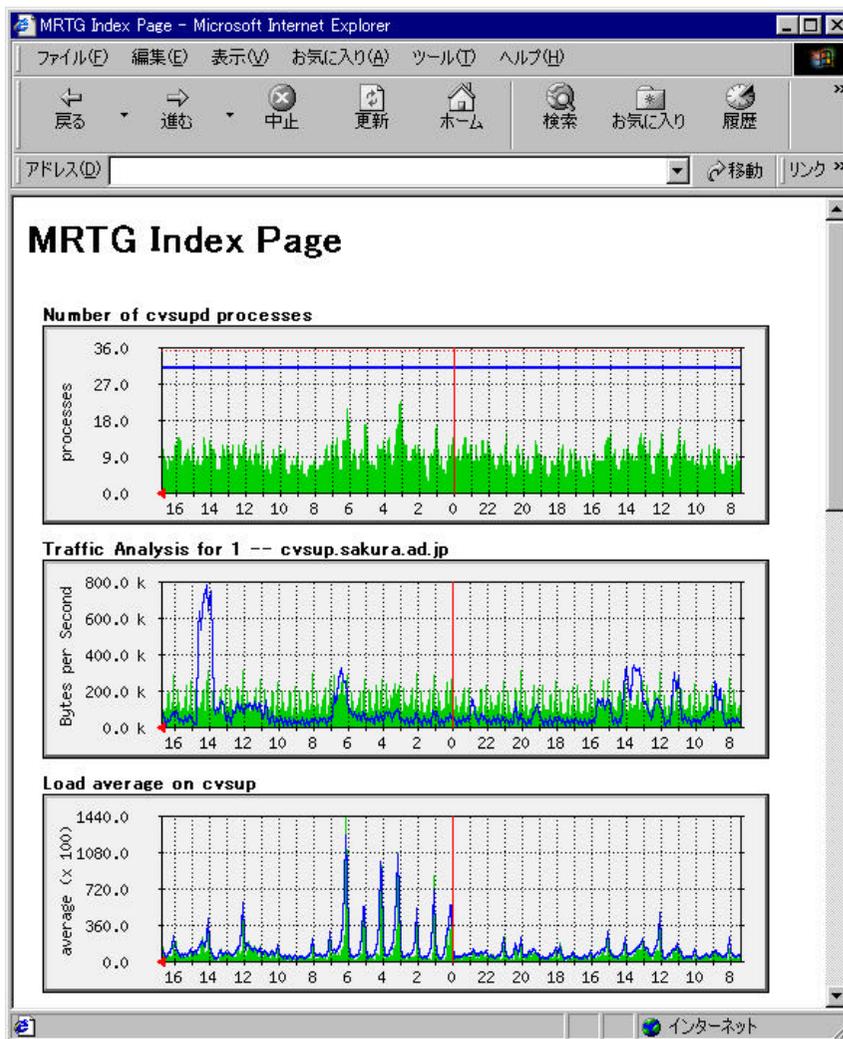
ただ、数百台レベルのサーバを同時にアップデートするとCVSupサーバへの負荷が心配

CVSupサーバ情報

10月からCVSup.jp.freebsd.org を運用中



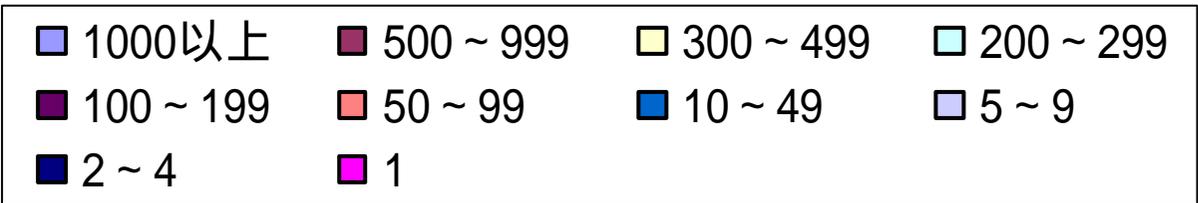
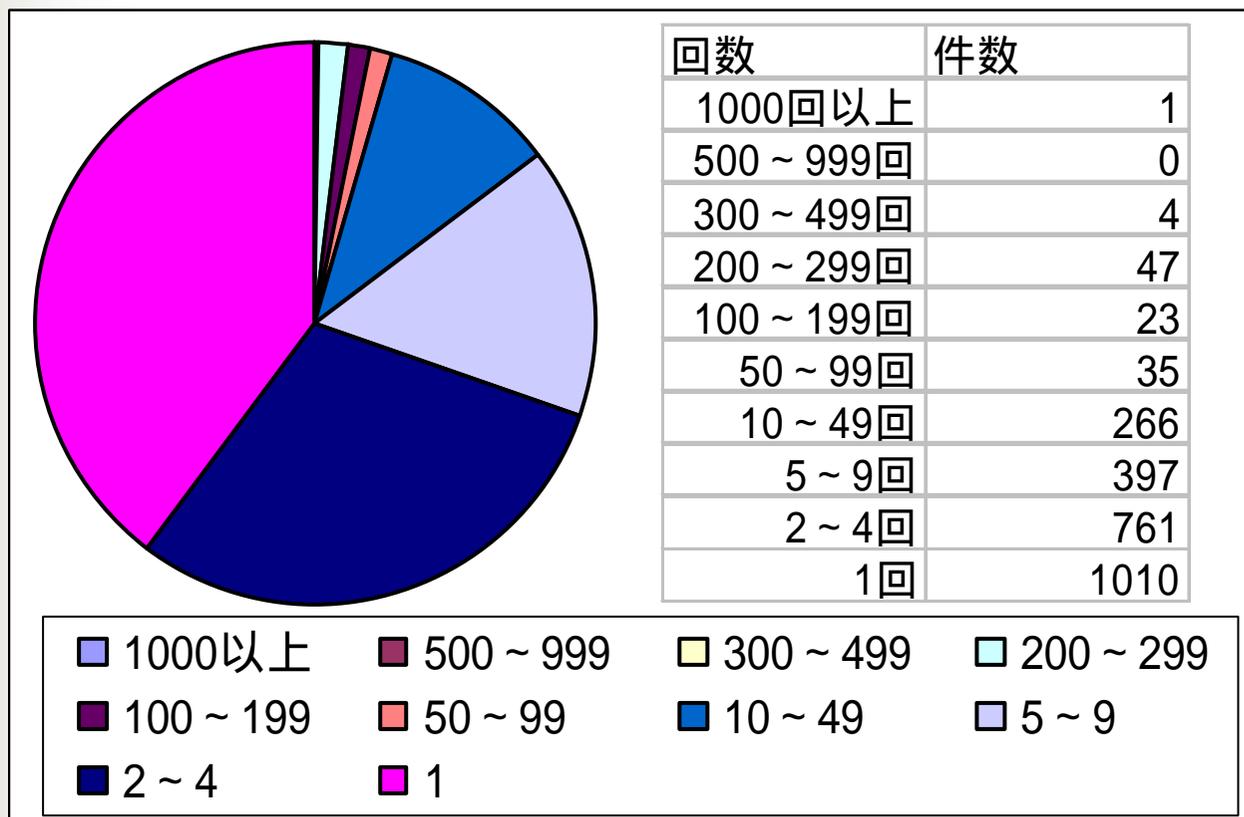
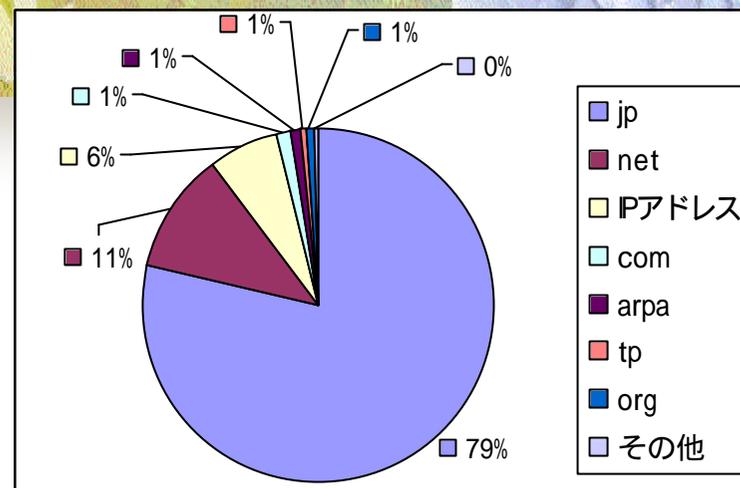
Memory	1.5GB
CPU	P3-866MHz × 2
Network	100Mbps
SCSICard	Adaptec 19160B



CVSupサーバ情報

2002/11/15から1週間の統計

全アクセス : 29149件
ユニークアクセス : 2544件





ご清聴有難うございました。