DNS Working Group 報告

第6回 ETJP 全体ミーティング
2004/9/14

藤原和典
株式会社日本レジストリサービス(JPRS)
fujiwara@jprs.co.jp

DNS-WG status

□現状報告

○DNS実装の評価が遅れています

○9月末から10月のETJP報告書に間に合わせる

○DNSSEC評価

▹未着手
▹将来的に評価

□今回の報告内容

○DNSSEC非対応の場合でのENUM実現可能性

DNS-WG:概要と目的

概要

- □日本国内で展開しうるENUMのDNSモデル
  - ○定義
  - ○要求仕様
  - ○評価基準
  - ○現在のDNS実装を性能評価
- □DNSSECのENUMへの適用について検討と評価

活動の成果

- □ENUM DNSに関するモデル・要求仕様
- □DNSサーバ評価結果
- □DNSSECのENUMへの適用についての調査結果

DNS-WG: 活動のマイルストーン:

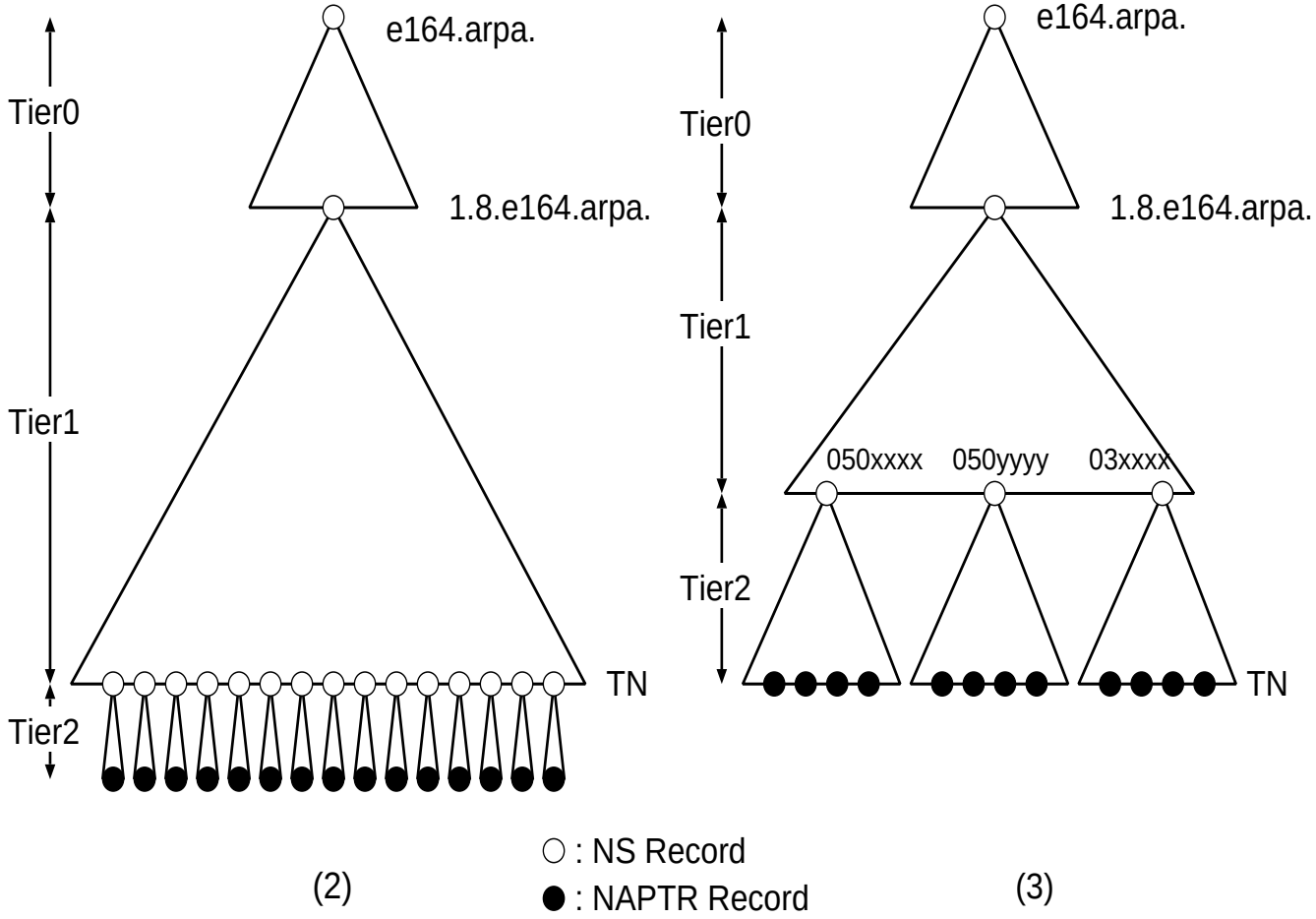
- □2004/3　DNSSEC対応レジストリシステムの提供
- □2004/4/E　モデル定義
- □2004/4/E　要求仕様策定
- □2004/5/E　評価基準策定
- □2004/6　DNSSECについての中間報告
- □2004/6/E　テスト環境構築
- □2004/9/E　DNSパフォーマンス評価報告
- □2004/9/E　DNSSECについての報告

DNS-WGで想定するENUM DNS要求条件

- □目的
  - ○トライアルおよび将来のENUMの商用利用に向けた基礎データ収集
  - ○DNSサーバの負荷について実機検証
- □DNS-WGで想定するENUMモデル
  - ○電話サービスのアドレス解決にENUMを用いる場合を想定
  - ○電話契約数、通話数をもとに必要条件を決定
- □登録数
  - ○系全体で1億件程度
- □DNSパフォーマンス条件
  - ○系全体で5万qps程度
- □DNSデータ更新頻度条件
  - ○変更までに30分以内

DNS-WG:想定するENUM DNS Tier構造

ENUM研究グループ報告書 23,24ページ (2)と(3)を検討



(2)たけのこモデル　　(3)割り当て単位モデル

DNS-WG: DNSサーバ性能評価

- □DNSコンテンツサーバ単体での能力測定

- □共通のテストデータ
  - ○たけのこモデル、割り当て単位モデル
  - ○Tier1、Tier2 それぞれ
  - ○パラメータで電話番号数を指定
    - ▹データが主記憶に入り切るサイズのデータ
    - ▹主記憶の量により、測定できるエントリ数が変わる
  - ○DNSSEC適用時についても調査(予定)

- □測定項目
  - ○DNSデータロード時間・リロード時間
  - ○メモリ使用量(プロセスサイズ)
  - ○サーバ問い合わせ応答性能(queryperfを使用)

- □その他テスト条件
  - ○できる範囲でそろえる

テストデータ(1)

□たけのこ Tier1

○機械的に生成するゾーンファイル
○エントリ数を指定して生成
○一つの番号あたり、NS RR 2行

```
9.8.7.6.5.4.3.2.1.0.1.8.e164.arpa. IN NS ns1.etjp.jp.
9.8.7.6.5.4.3.2.1.0.1.8.e164.arpa. IN NS ns2.etjp.jp.
```

□たけのこ Tier2

○機械的に生成する多数のゾーンファイル
○一つの番号にNAPTR RR 1行, NS RR 2行
○ゾーン数を指定して生成
○named.confも生成

```
$ORIGIN 9.8.7.6.5.4.3.2.1.0.1.8.e164.arpa.
  IN SOA
  IN NS ns1.etjp.jp.
  IN NS ns2.etjp.jp.
  IN NAPTR 100 0 "u" "E2U+sip" "!^.*$!sip:810123456789@example.jp!" .
```

テストデータ(2)

- □局番単位 Tier1
  - ○機械的に生成するゾーンファイル
  - ○局番数を指定、15万エントリ程度
  - ○一つの局番あたり、NS RR 2行

```
$ORIGIN 1.8.e164.arpa.
@ IN SOA ns0.etjp.jp. postmaster.etjp.jp. (1 1H 5M 7D 10M)
 IN NS ns1.etjp.jp.
 IN NS ns2.etjp.jp.
0.0.0.0.0.0 IN NS ns1.isp000.jp.
0.0.0.0.0.0 IN NS ns2.isp000.jp.
1.0.0.0.0.0 IN NS ns1.isp001.jp.
...
```

- □局番単位 Tier2
  - ○機械的に生成する多数のゾーンファイルとnamed.conf
  - ○各ゾーンファイルに1万番号
    - ▹1番号ごとにNAPTR RR 1行
  - ○ゾーン数を指定して生成

```
$ORIGIN 0.0.0.0.0.0.1.8.e164.arpa.
$TTL 120
@ IN SOA ns1.etjp.jp. postmaster.etjp.jp. (1 1H 5M 7D 10M)
 IN NS ns1.etjp.jp.
 IN NS ns2.etjp.jp.
0.0.0.0 IN NAPTR 100 0 "u" "E2U+sip" "!^.*$!sip:00000000@sipisp.jp!" .
1.0.0.0 IN NAPTR 100 0 "u" "E2U+sip" "!^.*$!sip:00000001@sipisp.jp!" .
2.0.0.0 IN NAPTR 100 0 "u" "E2U+sip" "!^.*$!sip:00000002@sipisp.jp!" .
...
```

DNS-WG: 評価基準

□要件を満たす実装と、その仕組み(組合せ)の提案を目指す

- ○各要素を実用的に実現する方法
- ○DNSコンテンツサーバ単体での性能測定
- ○コンテンツサーバを何台に分割すると要求を満たせるか
- ○実用台数組み合わせて実現できるモデル
  - ▹運用可能な実用的なシステム数を10から100と想定

DNS-WG: 調査対象DNSサーバ

- BIND8: 8.3.7
  - ISCが開発、従来の標準
  - reload中に無応答になる
  - Authoritative server機能とFull Resolver機能を持つ
- BIND9: 9.3.0rc4
  - ISCが参照実装として開発
  - DNSSEC対応
  - Authoritative server機能とFull Resolver機能を持つ
- NSD: 2.1.1
  - NL NetLabsが開発
  - DNSSEC対応
  - Authoritative server機能のみ
- djbdns: 1.0.5
  - D. J. Bernstein助教授が開発
  - Authoritative server機能とFull Resolver機能を分離
  - データセットの切替えは瞬時に可能

DNS-WG: 測定環境

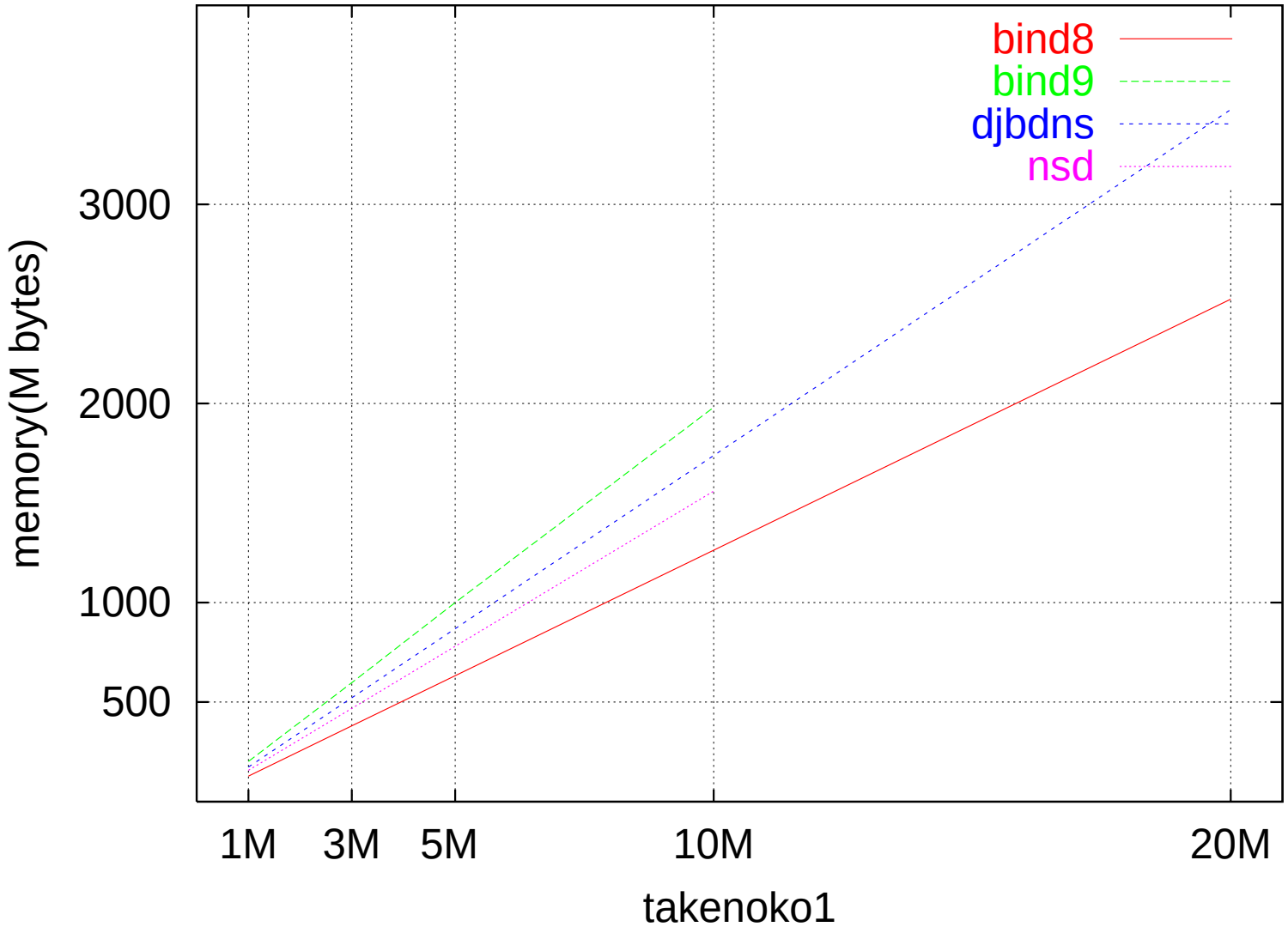
□DNSサーバPC-------Ethernet-------測定用PC

- ○測定対象DNSサーバPC 1台
- ○測定用PC 1台
- ○同一ネットワークに接続(100baseTX hub)

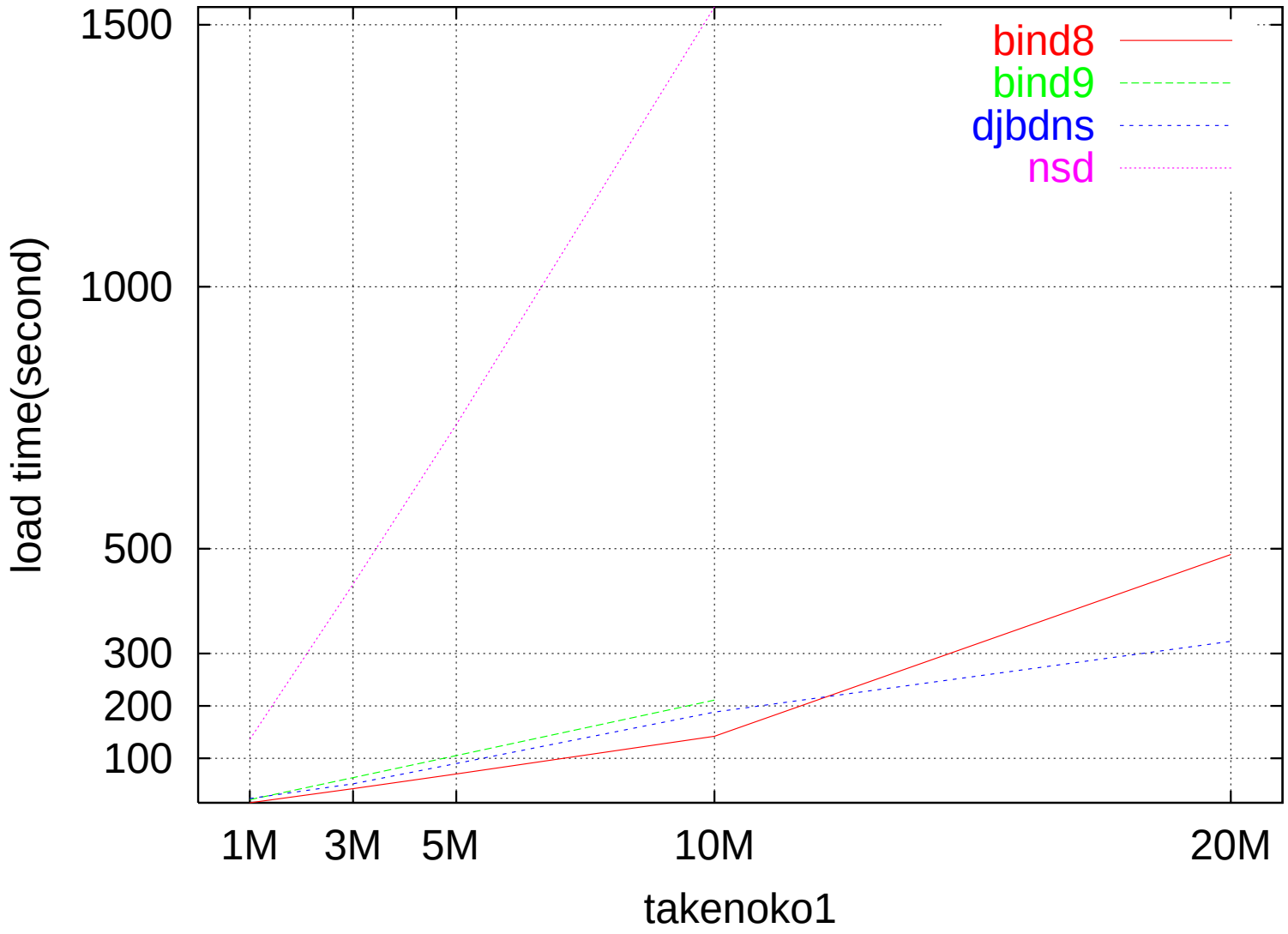
□テスト環境例

- ○DNSサーバPC
  - ▹Pentium4-3GHz, memory 2.5GB, DragonFlyBSD 1.1-CURRENT
  - ▹Pentium4-3GHz, memory 1.5GB, FreeBSD 4.10-RELEASE
- ○queryperf PC
  - ▹Pentium4-2.7GHz, FreeBSD
- ○ethernet
  - ▹100baseTX HUB

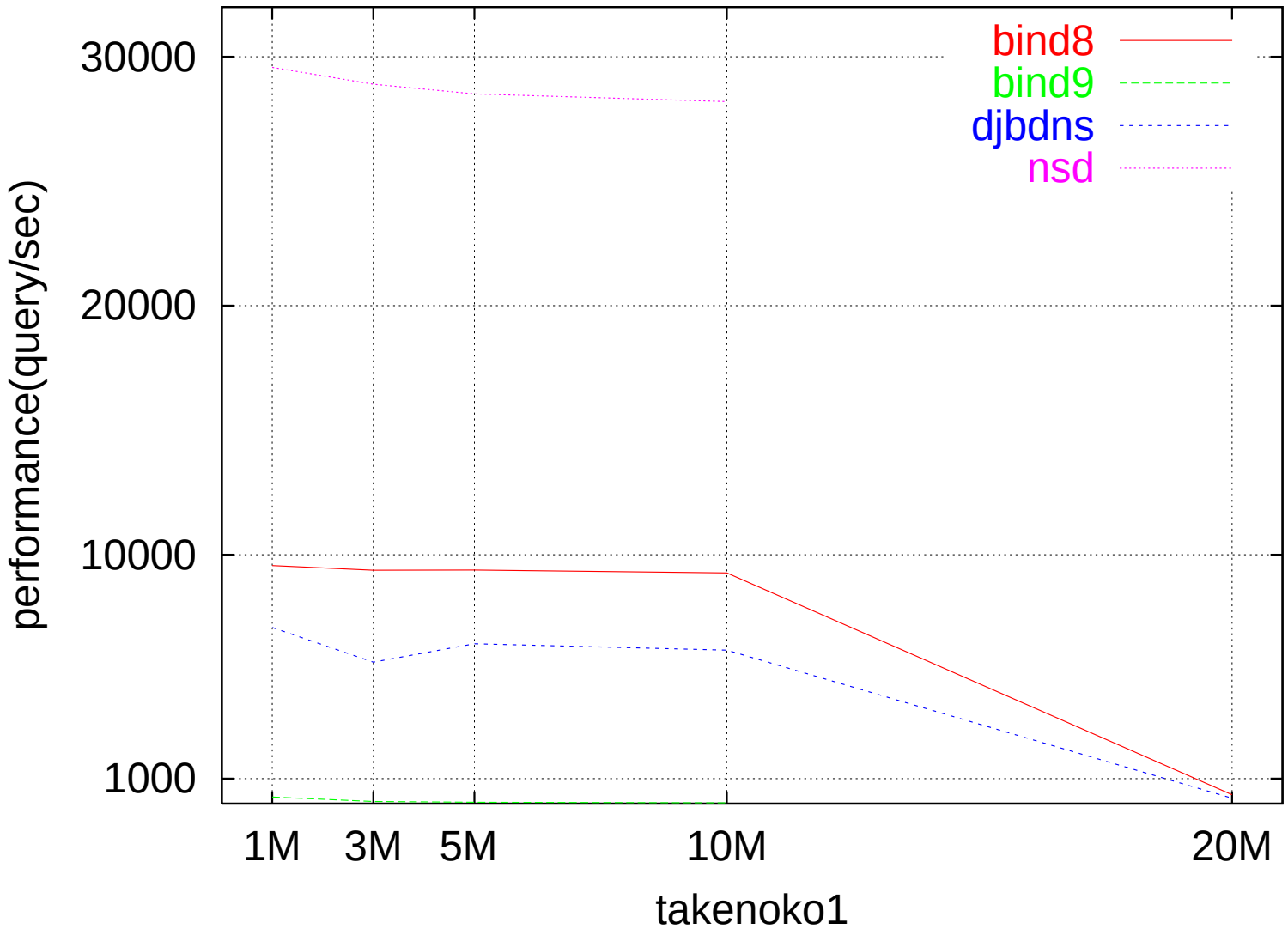
主記憶2.5G,たけのこ1 メモリ使用量



主記憶2.5G,たけのこ1 ロード時間



主記憶2.5G,たけのこ1 応答性能



たけのこモデルTier1 評価結果(1)

- □OS種別による大きな違いは見受けられない

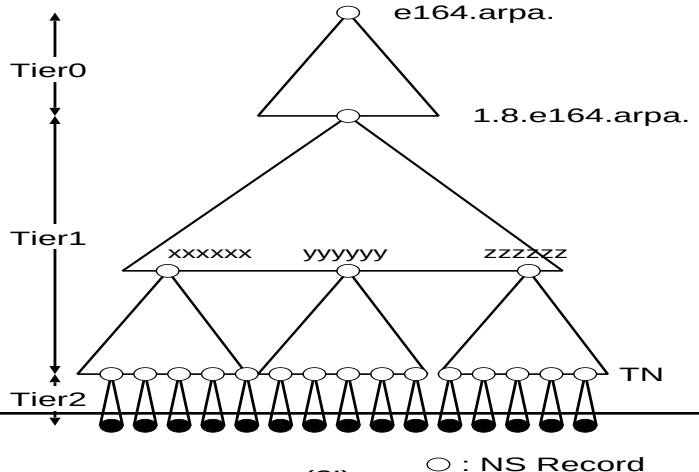
- □登録数
  - ○どのサーバでも 1000万番号 2000万エントリまでは正常動作
    - ▹10システムにわけることで1億対応可

- □サーバパフォーマンス
  - ○NSDでは 1000万エントリで27000qpsを処理可能
  - ○DNSコンテンツサーバを2台用いることで5万qps程度まで処理可能
  - ○BIND8, djbdnsの場合も10台程度並列に用いることで5万qps程度まで対応可
  - ○BIND9では100qps以下

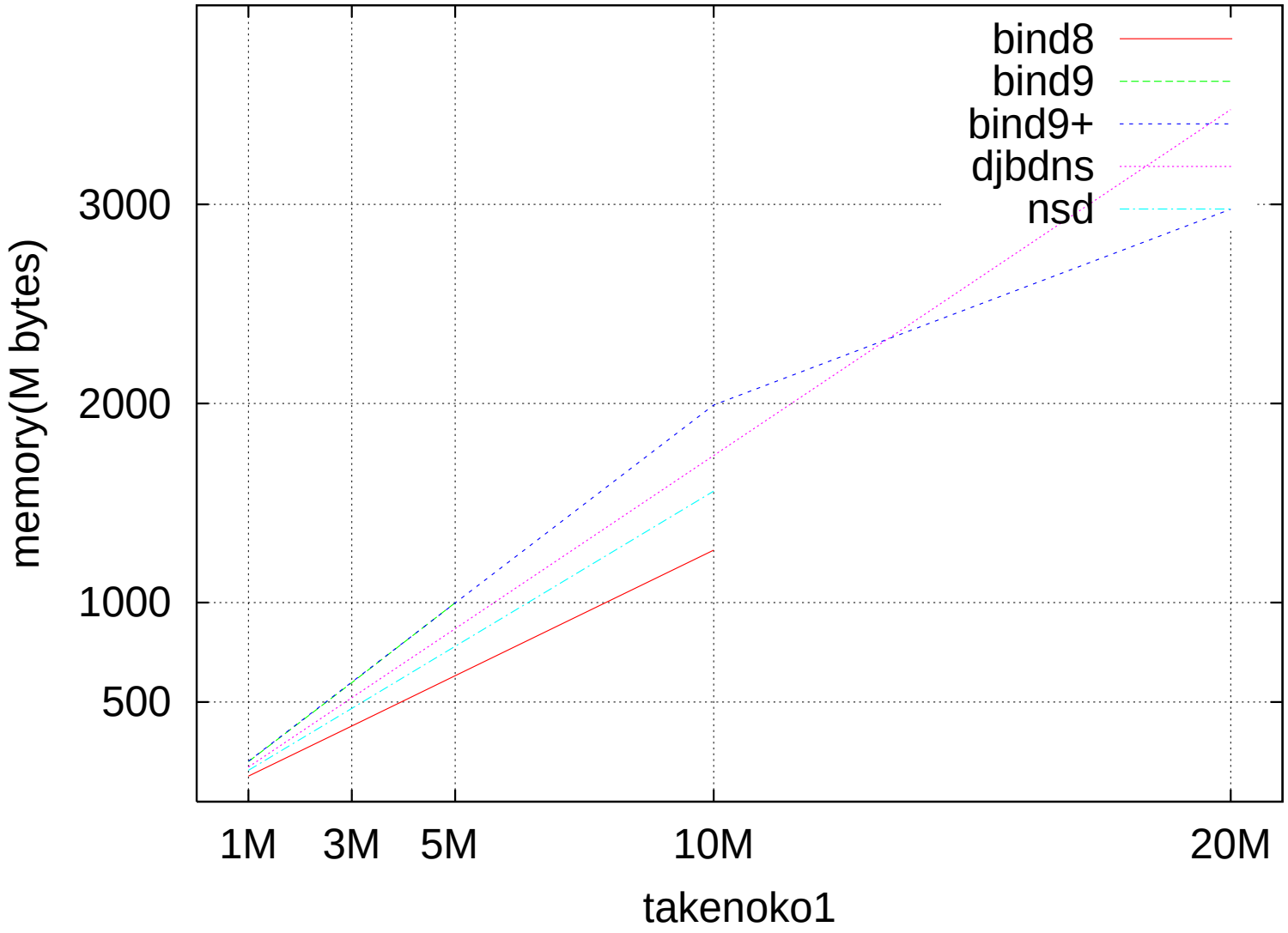
- □更新頻度条件
  - ○NSDの場合はデータのコンパイル・ロードだけで30分弱かかる
  - ○BIND 8, djbdnsの場合は4分弱でロード可能

たけのこモデルTier1 BIND9改善案

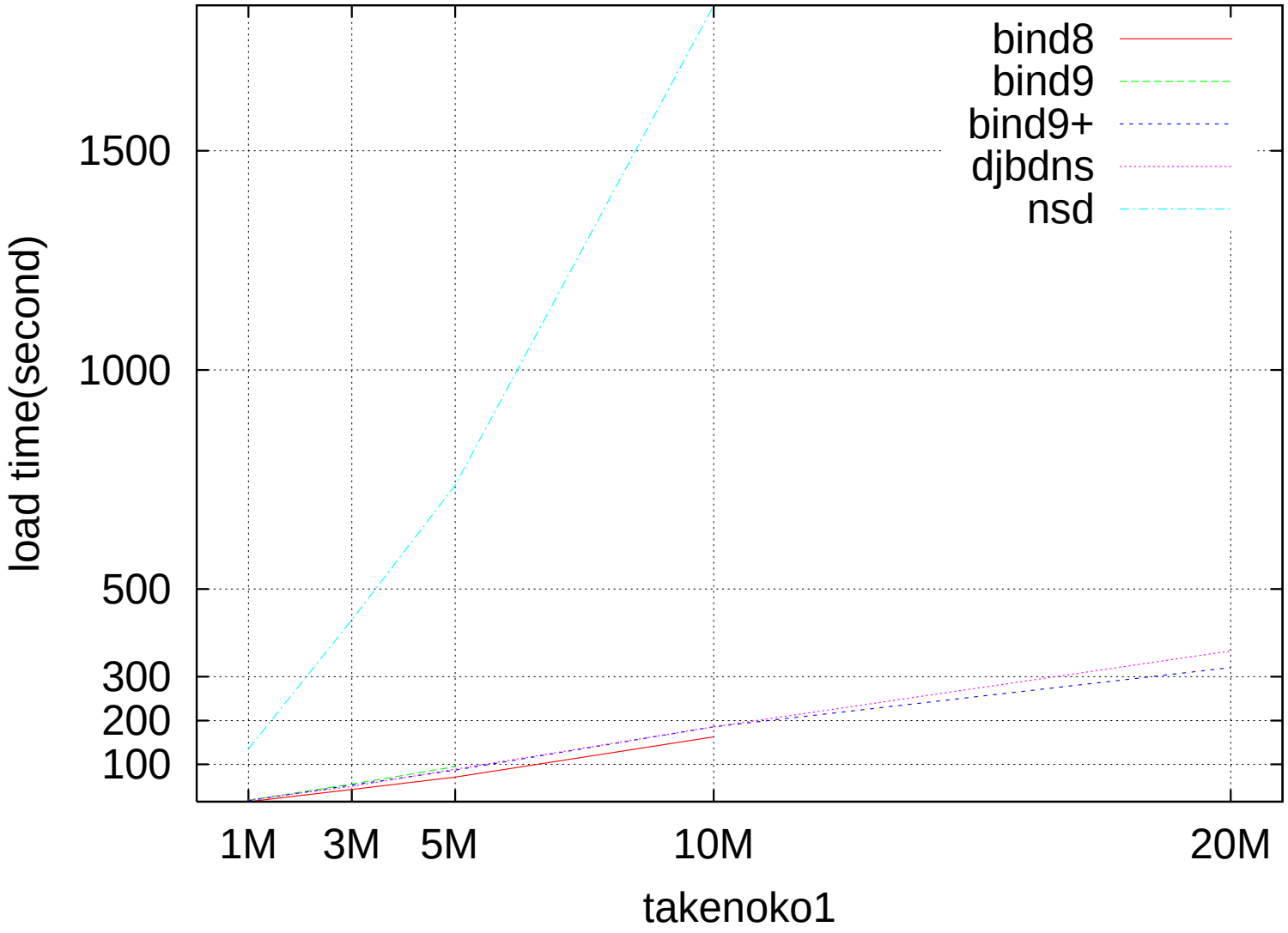
- □遅い原因は？
  - ○BIND9の内部構造と推測
  - ○ISC開発担当者に質問
    - ▹階層が深いENUMゾーンを持たせると性能が出ない
  - ○回答
    - ▹BIND9の内部構造に起因
    - ▹4桁程度で階層を分割して一台に持たせればよい(予想通り)
- □BIND9向けゾーンファイルによる再測定
  - ○8桁の1000万エントリゾーン
  - ○→下位4桁10000エントリのゾーン1000個に分割
  - ○一台にすべてのゾーンを持たせる
  - ○以下bind9+とする



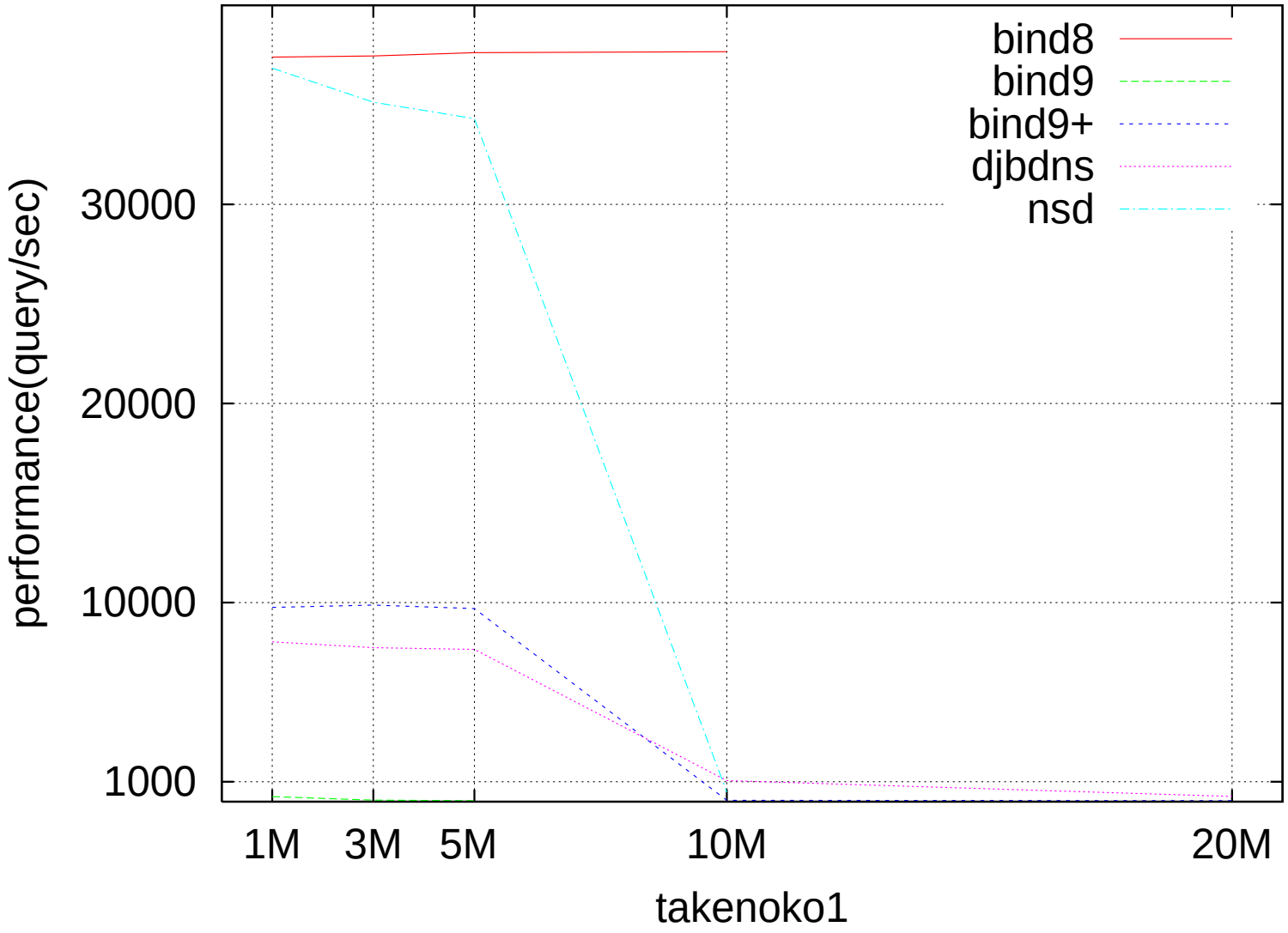
主記憶1.5G,たけのこ1 メモリ使用量



主記憶1.5G,たけのこ1 ロード時間



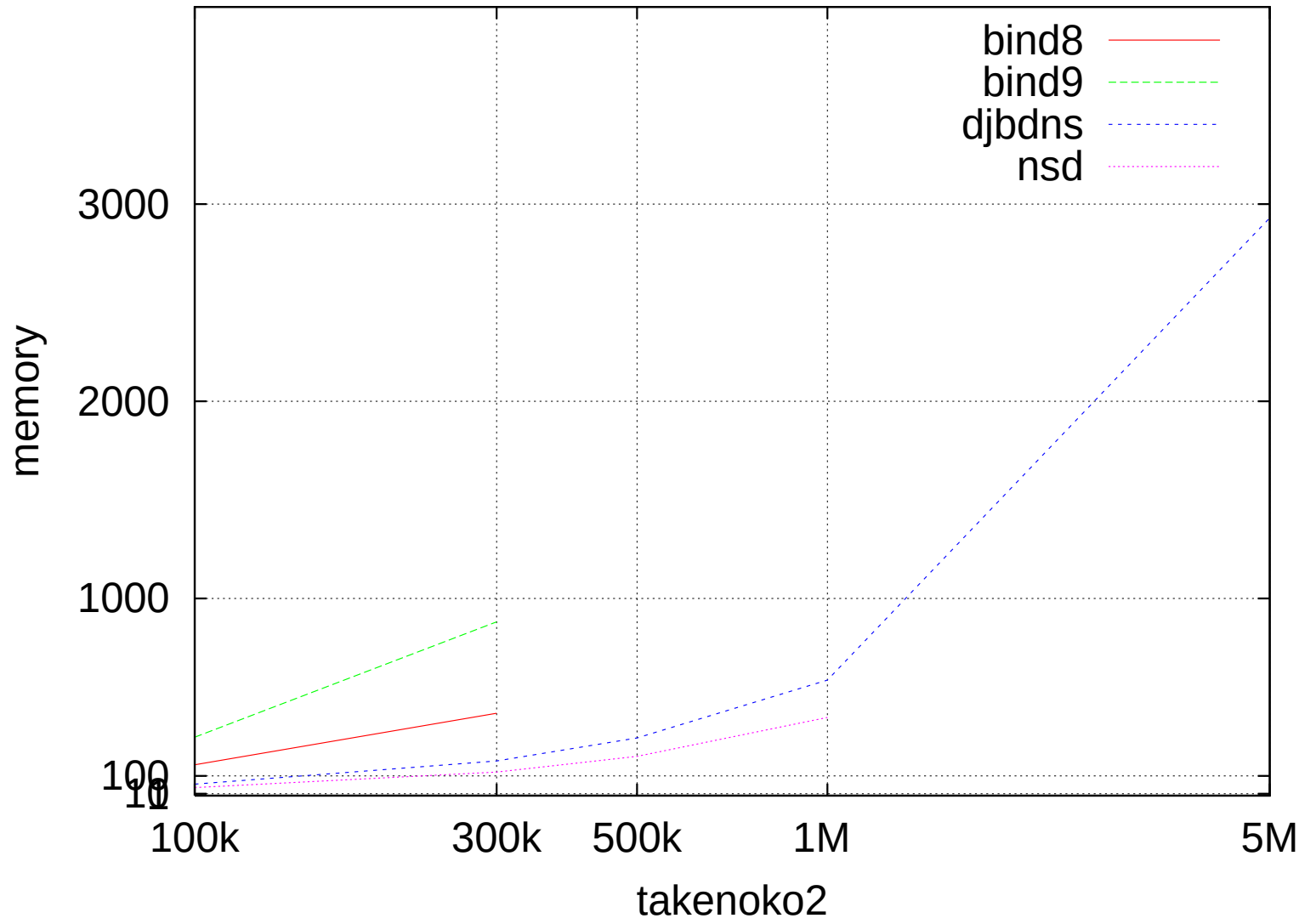
主記憶1.5G,たけのこ1 応答性能



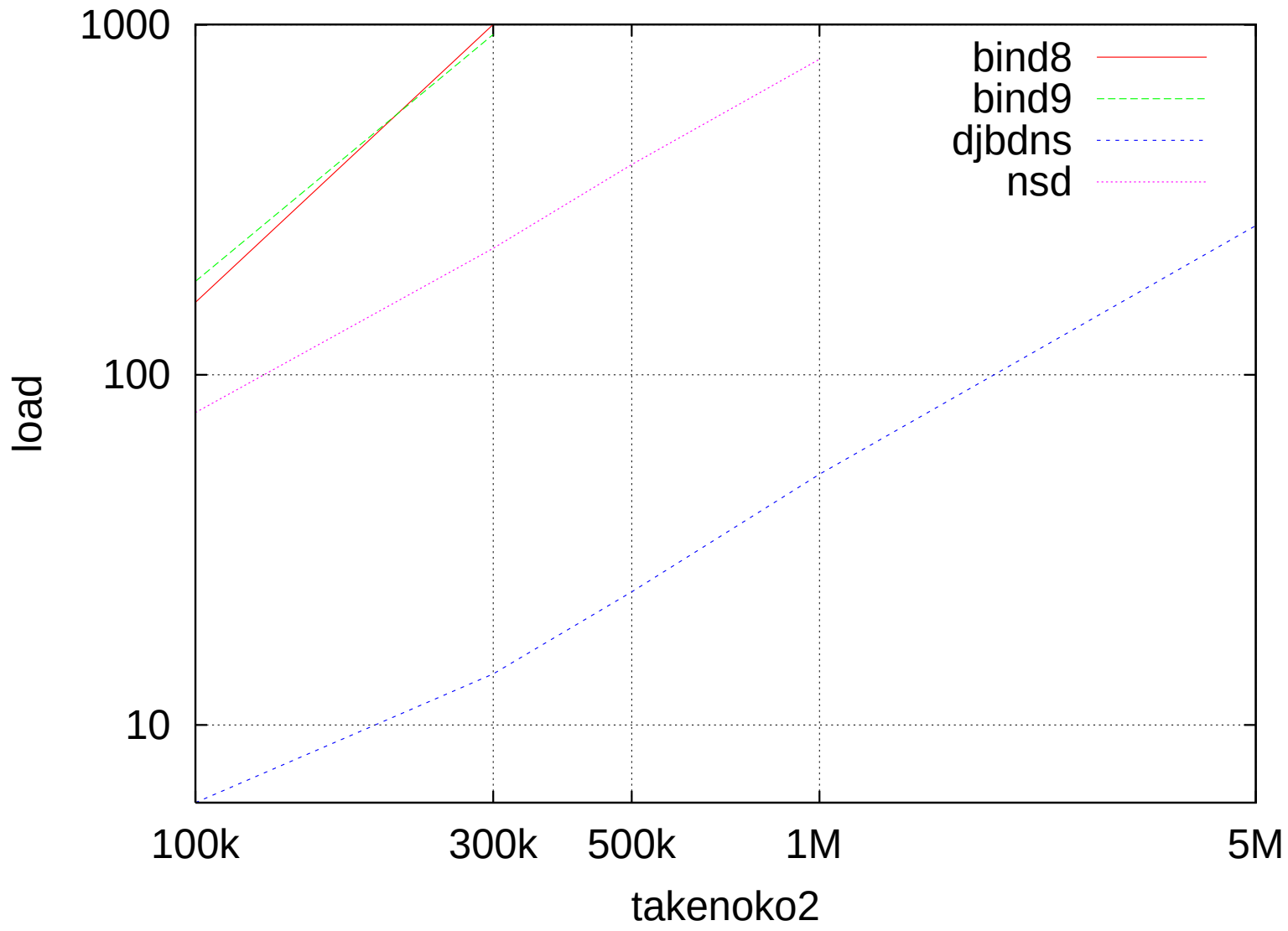
たけのこモデルTier1 評価結果(2)

- BIND 9 の場合も、BIND 9に適したデータを作れば、1万qps程度の性能が出る
- メモリ使用量、読み込み時間は変化しない
- 今回の評価では主記憶が少ない分だけ、扱えるデータが減っている
  - 2.5G → 1.5G (OSが使う部分も含む)
  - 今回は、プロセスが使えるメモリはほぼ半分になっている
  - その結果、今回の主記憶量では、500万エントリまで

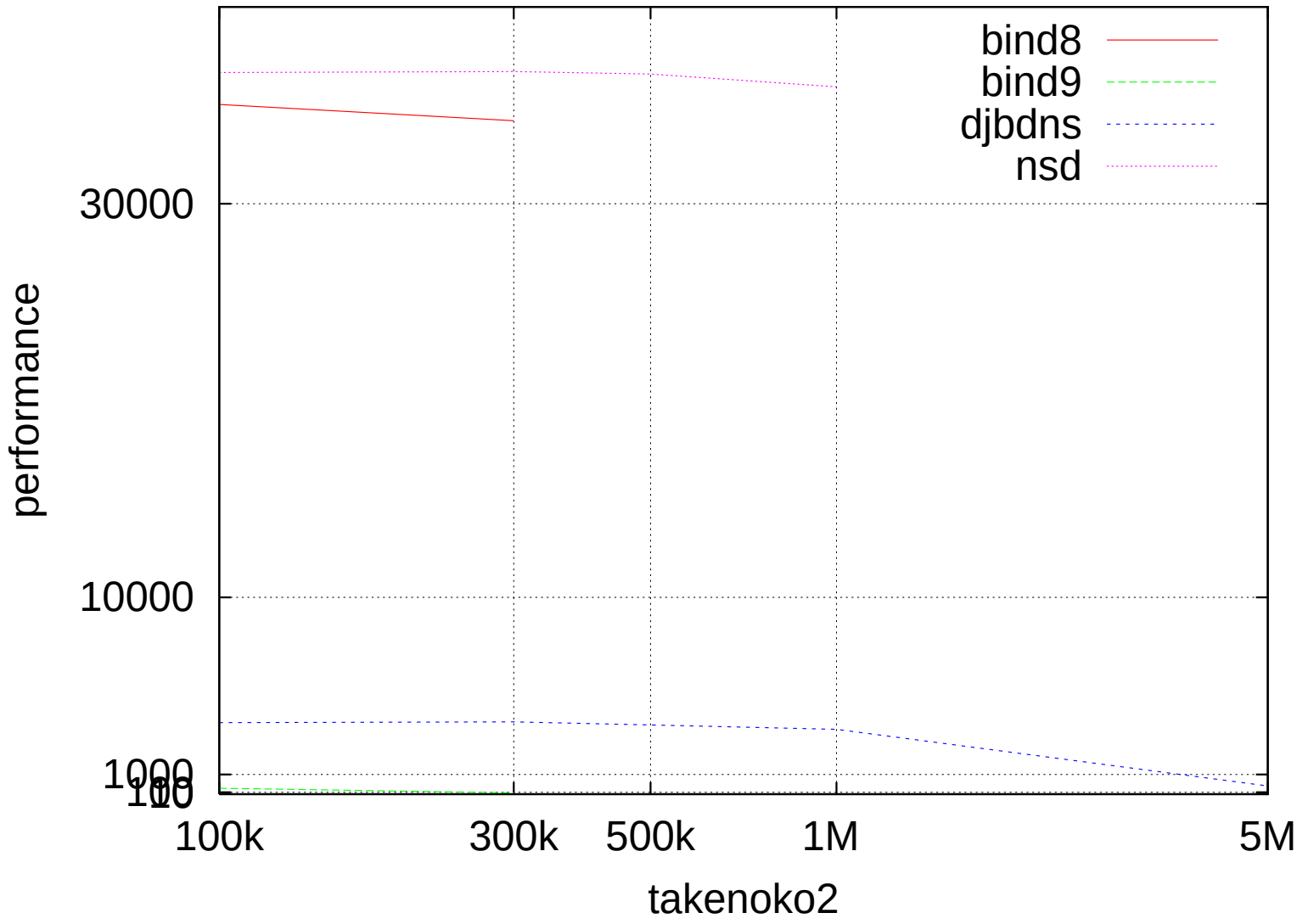
主記憶1.5G,たけのこ2 メモリ使用量



主記憶1.5G,たけのこ2 ロード時間



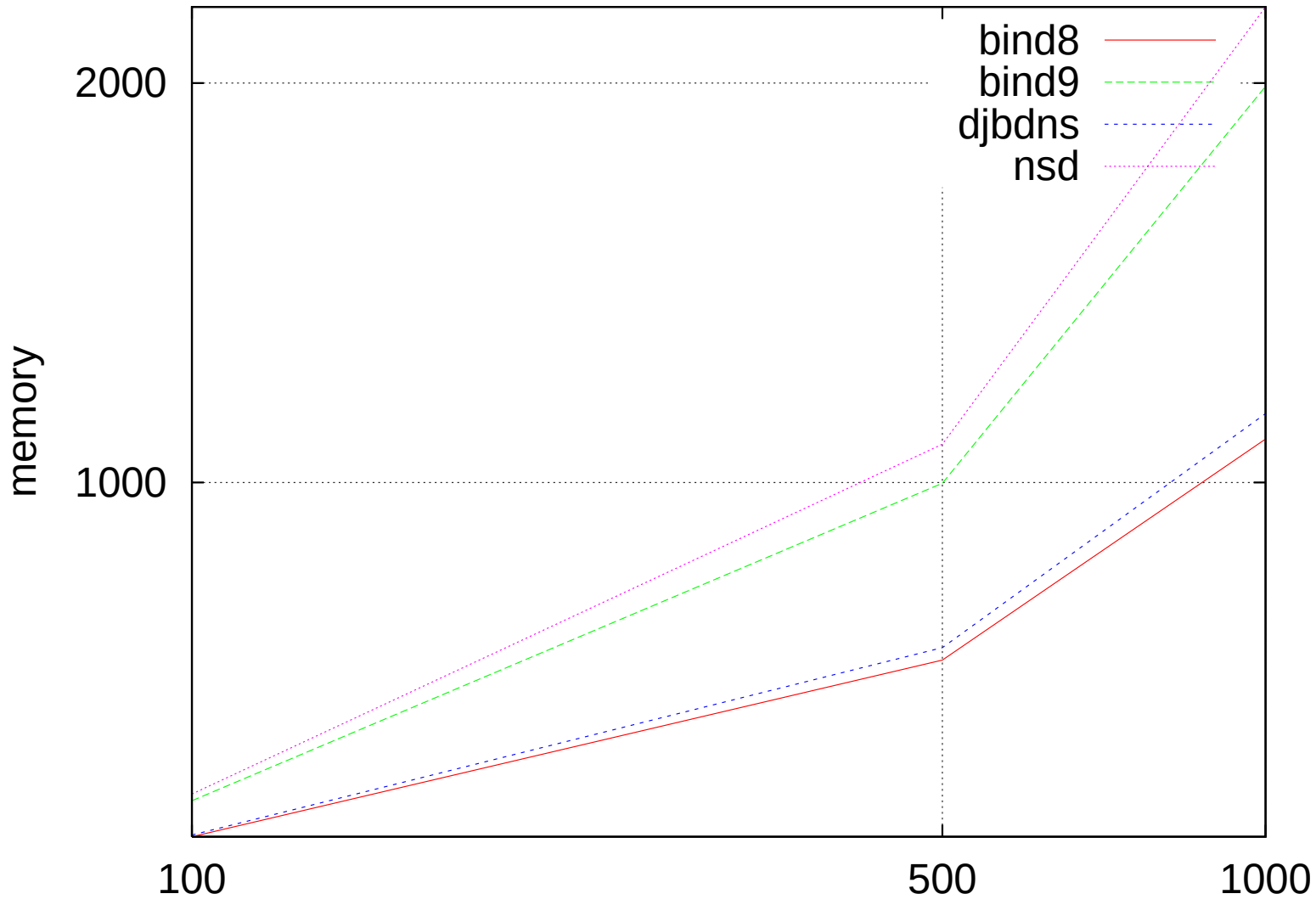
主記憶1.5G,たけのこ2 応答性能



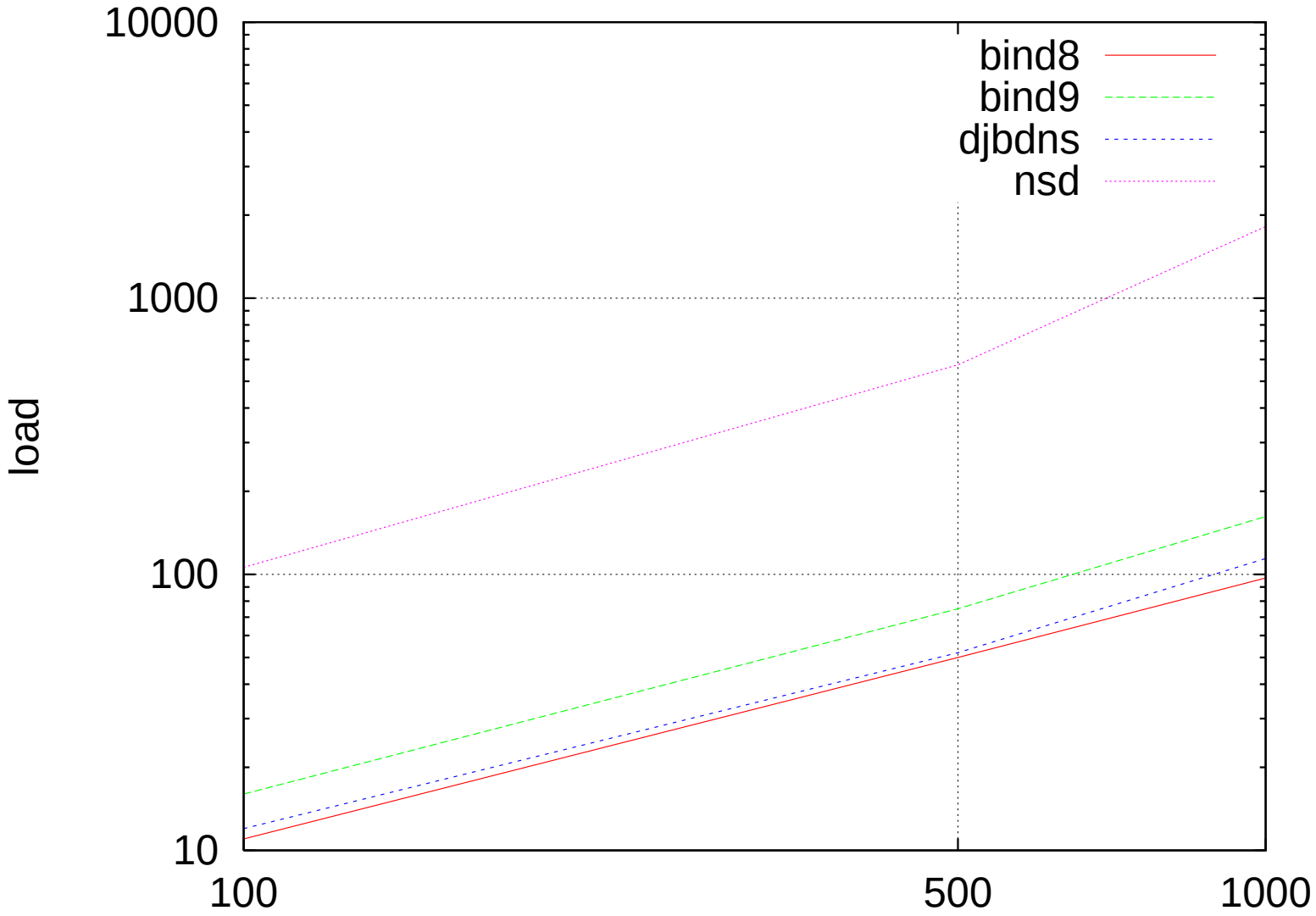
たけのこモデルTier2 評価

- 登録数(保持可能数)
  - BIND 8, BIND 9では30万ドメイン
  - NSDでは100万ドメイン
  - djbdnsでは500万ドメイン
- サーバパフォーマンス
  - NSD、BIND 8は3万qps以上の性能を示した
  - BIND 9は性能がでない
- 更新頻度条件
  - BIND8, 9は30万ゾーンの読み込みに20分程度の時間がかかる
  - djbdnsのデータファイル変換の時間がもっとも小さい
    - 500万件でも4分程度

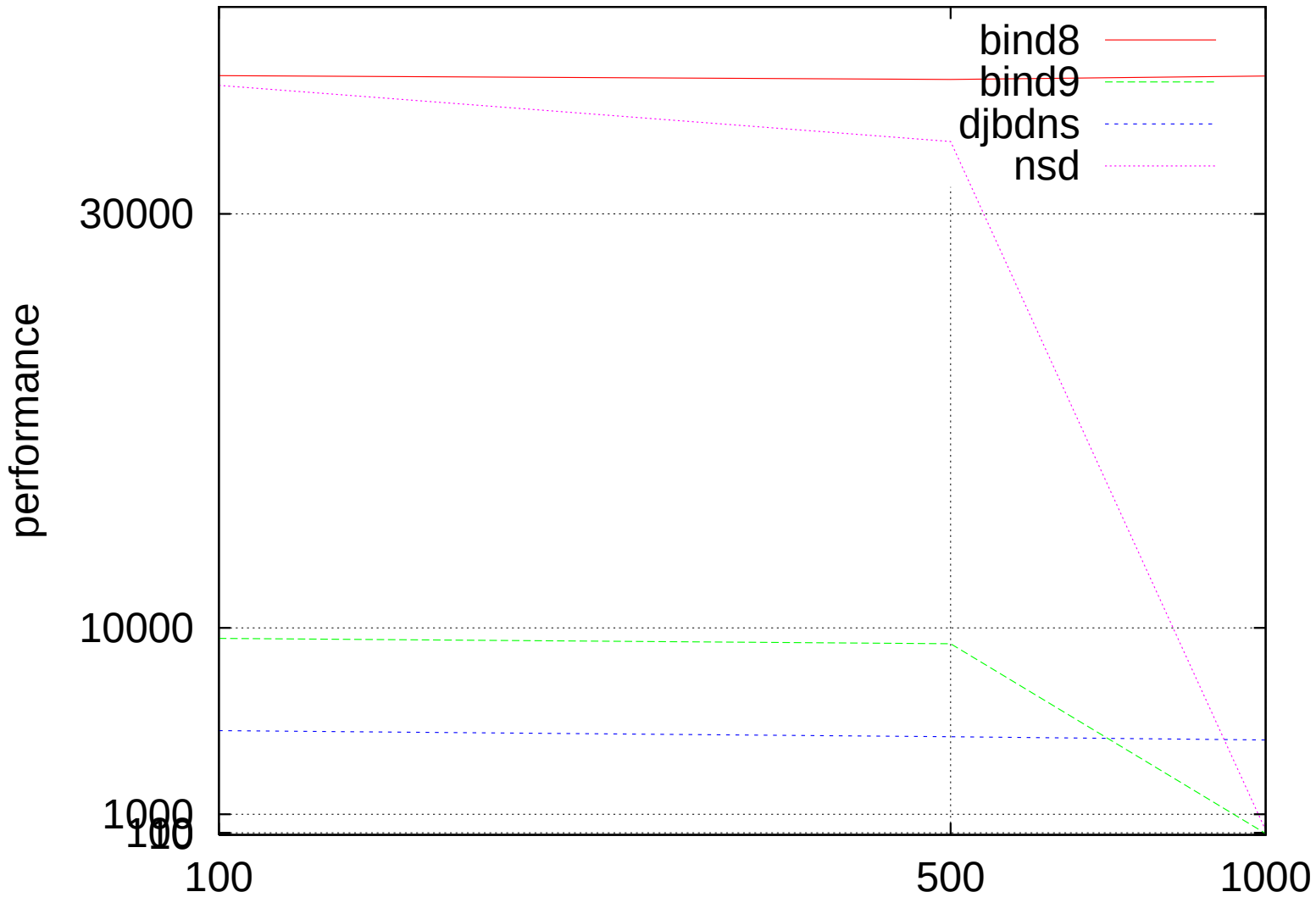
主記憶1.5G,割当単位2 メモリ使用量



主記憶1.5G,割当単位2 ロード時間



主記憶1.5G,割当単位2 応答性能



割り当て単位モデルTier2 評価

□登録数

○どのサーバでも1万エントリのゾーンを1000ゾーン蓄積できた

□サーバパフォーマンス

○BIND8は1000ゾーンまで3万qps以上の性能を示した

○NSD, BIND9は500ゾーンまで(メモリの制約)

○djbdnsはほぼ一定であるが、他に比べ低い

□更新頻度条件

○BIND8, 9は比較的短時間で読み込むことができる

○djbdnsのデータファイル変換の時間が小さい

○NSDはゾーンのコンパイル、読み込みに他の実装の10倍程度の時間がかかる

まとめ(1)

- 32bit CPU/OSの性能測定となってしまった
  - 1000から2000万エントリでは32bit CPU/OSのメモリ空間を使い切る
- 報告書には、モデルごとに可能な組合せを記述する予定
- データ更新時の性能については別途評価中

まとめ(2)

□djbdns

○メモリ空間ぎりぎりまで使わなければ、安定して4000～6000qps程度

○短めのデータ変換時間

○ロード時間ゼロ(起動時にメモリにデータを展開しない)

□NSD

○よいパフォーマンス

○長いデータ準備・ロード時間

□BIND9

○不得意なデータ構造あり(パフォーマンスが出ない)

□BIND8

○DNSSECなしではよい選択肢である

○読み込み中は問い合わせに答えない(運用上の考慮が必要)

DNS-WG: 予定

- 2004年9月まで 測定
- 2004年9月末から10月 評価結果報告