PostgreSQL Birds-of-a-Feather

Josh Berkus, Mark Wong

LinuxWorld Expo

August 4, 2004

©Copyright 2004 OSDL, All rights reserved.

Topics

- Introduction
  - OSDL Mission
  - What Mark does at the OSDL
- Using DBT-2 to evaluate separating data, logs, and indexes
- Using DBT-3 to test query optimizer changes
- What does this do for the Linux community?

©Copyright 2004 OSDL, All rights reserved.

OSDL Mission

- To be the recognized center of gravity for Linux; the central body dedicated to accelerating the use of Linux for enterprise computing through:
  - Enterprise-class testing and other technical support for the Linux development community.
  - Marshalling of Linux-industry resources to focus investment on areas of greatest need thereby eliminating inhibitors to growth.
  - Practical guidance to our members - vendors and end users alike - on working effectively with the Linux development community.

©Copyright 2004 OSDL, All rights reserved.

What I Do

- Develop real-world database workloads to test system performance
  - DBT-1 Web Commerce
  - DBT-2 OLTP
  - DBT-3 Decision Support

©Copyright 2004 OSDL, All rights reserved.

DBT-2

- OLTP workload
- Sequential writes to the log
- Random reads and write to data
- Gain some efficiency by separating the logs from the data

©Copyright 2004 OSDL, All rights reserved.

Starting Out

- PostgreSQL 7.4
- System Details
  - 4 x 1.5GHz Itanium 2
  - 16 GB RAM
  - 42 x 15K RPM drives
  - 3 SCSI Controllers, 6 channels
- Data, log and indexes are on a single partition using LVM

©Copyright 2004 OSDL, All rights reserved.

1st Step

- Separate logs from the table data and indexes
- Problem: PostgreSQL 7.4 doesn't support this
- Solutions:
  - Move the database log to another location and symlink (easy)
  - Query the database for the table identifiers, move the tables to another location, and symlink (harder)
  - Wait for 7.5 tablespace support

©Copyright 2004 OSDL, All rights reserved.

Results of reorganizing

- 7.5 Baseline (spread uniformly)

  Average Throughput: 4411 NOTPM
- Next step: determine best physical layout of disks using the tablespace feature
  - 7.5 Separating log from data

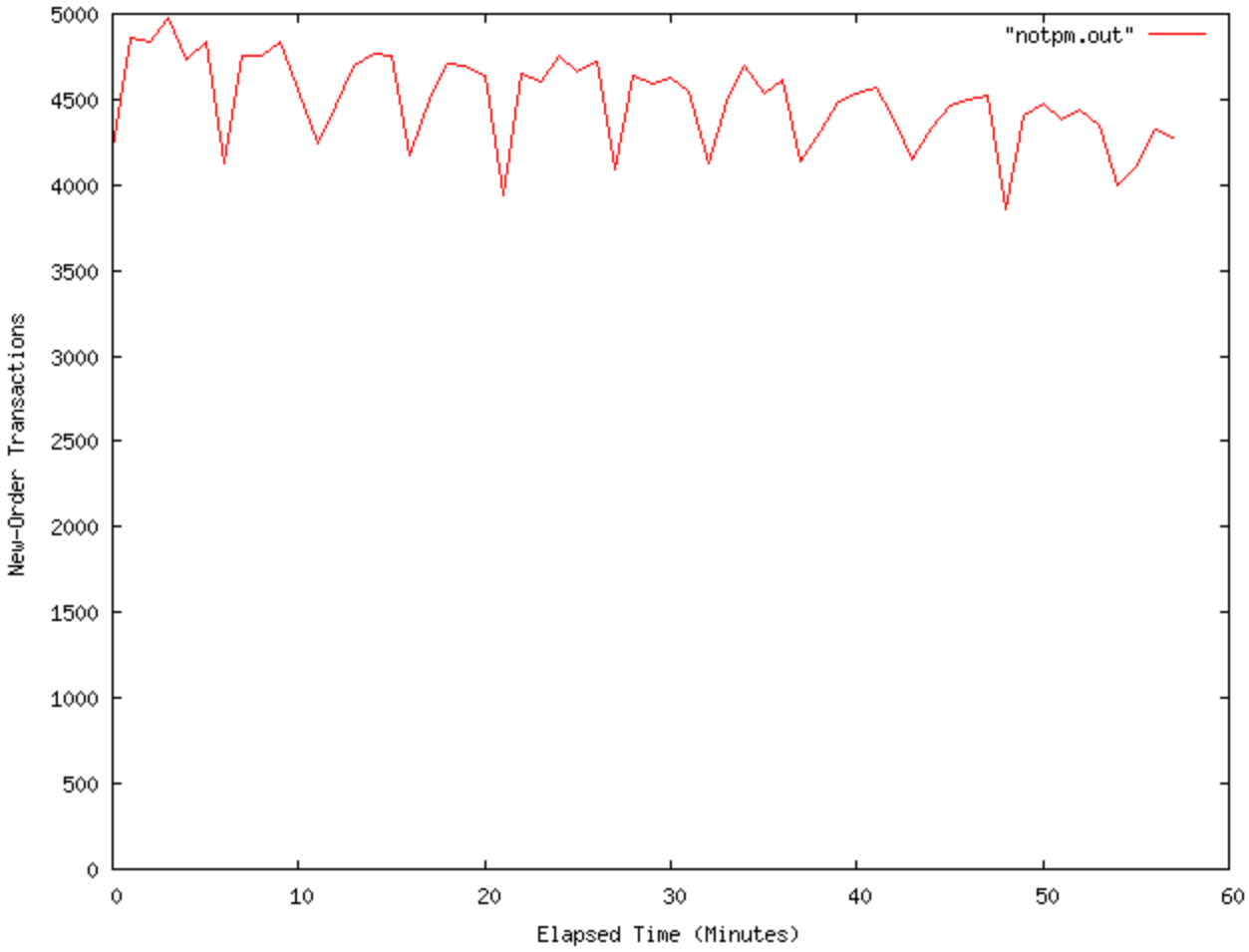
    Average Throughput:
  - 7.5 Every table/index or log on separate volumes

    Average Throughput: 4996 NOTPM

    13% improvement
  - Trying other layouts.

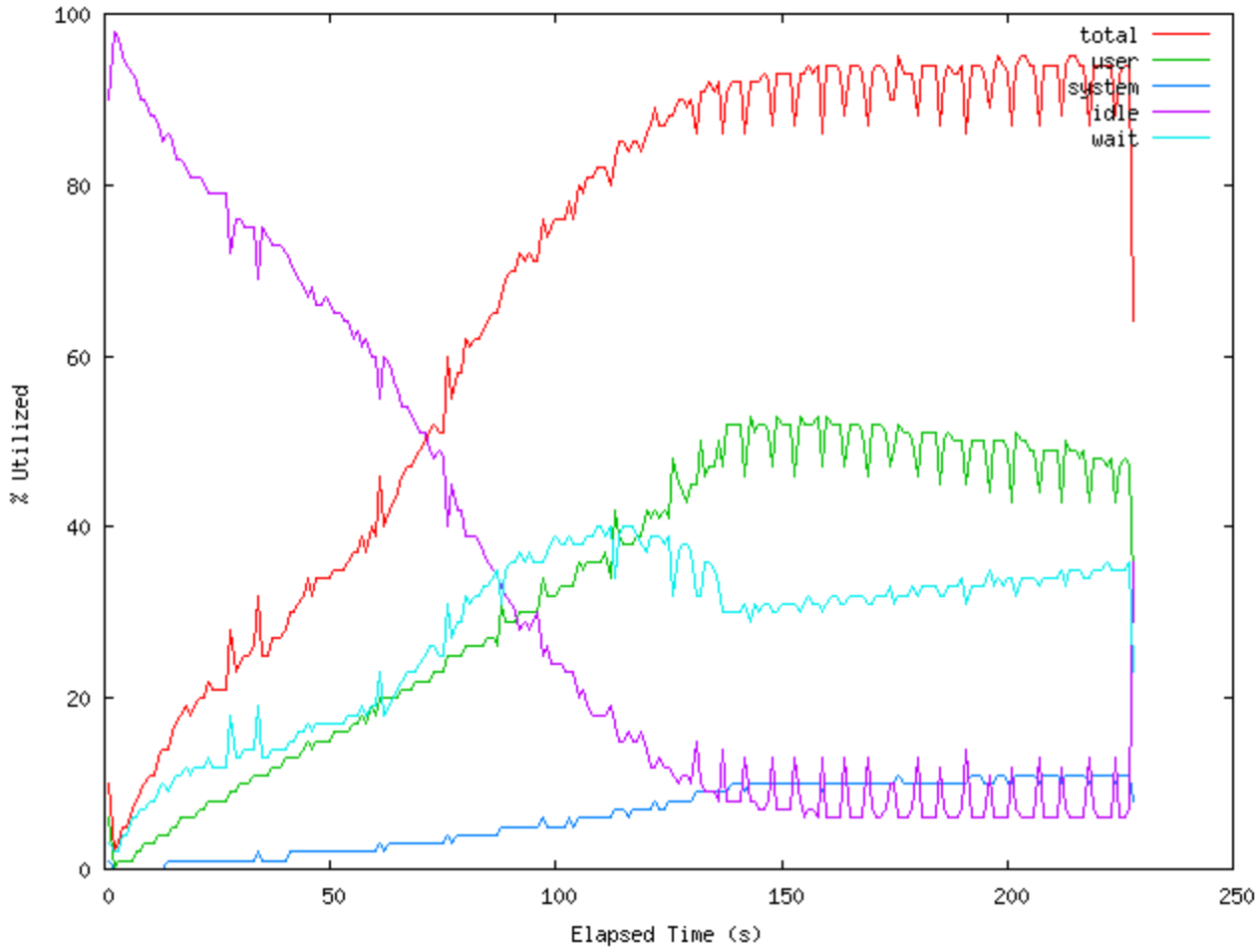
©Copyright 2004 OSDL, All rights reserved.

Throughput



©Copyright 2004 OSDL, All rights reserved.

CPU Utilization



©Copyright 2004 OSDL, All rights reserved.

Other performance issues encountered scaling to 4 way systems:

- Spinlock for IA32 systems too low (100 versus 1000)
  - Faster your CPU speed, the greater the improvement
- Rep:nop added to the end of the loop for performance
- Effects of Linux 2.6 I/O scheduler:

Io scheduler	File system	CPU	NOTPM	% improvement	Runid
Anticipatory	JFS	4	1642.0	Base	292768
Deadline	JFS	4	1725.5	5.09	291289
Noop	JFS	4	1659.85	1.09	292769

©Copyright 2004 OSDL, All rights reserved.

Other performance issues (cont)

- File System performance

File system	CPU	NOTPM	%relative to ext2
Ext2	4	1749.8	Base
Ext3	4	1176.95	-32.74
Xfs	4	1611.75	-7.89
Reiserfs	4	1688.35	-3.51
JFS	4	1725.5	-1.39

©Copyright 2004 OSDL, All rights reserved.

DBT-3

- Decision Support Workload
- Mostly reads
- Complex queries had to be manually rewritten because the optimizer wasn't able to generate a plan that executed within a reasonable amount of time

©Copyright 2004 OSDL, All rights reserved.

Query Planner

- http://archives.postgresql.org/pgsql-hackers/2003-12/msg00138.php
- "Currently, queries with complex "or group" criteria get devolved by the planner into canonical and-or filters resulting in very poor execution on large data sets." - Josh Berkus

©Copyright 2004 OSDL, All rights reserved.

General Example

```
SELECT t1.a, t2.b
FROM t1, t2
WHERE t1.a = t2.a
AND (
      ( t1.c = x
        AND t1.f IN (m, n, o)
        AND t2.d = v
        AND t2.e BETWEEN j AND k
      )
      OR
      ( t1.c = y
        AND t1.f IN (n, o, p)
        AND t2.d = v
        AND t2.e BETWEEN k AND h
      )
      OR
      ( t1.c = z
        AND t1.f IN (p, q)
        AND t2.d = w
        AND t2.e BETWEEN k AND h
      )
)
```

©Copyright 2004 OSDL, All rights reserved.

Teach the planner to do this

```
SELECT t1.a, t2.b
FROM t1, t2
WHERE t1.a = t2.a
AND t1.c = x
  AND t1.f IN (m, n, o)
  AND t2.d = v
  AND t2.e BETWEEN j AND k
UNION
SELECT t1.a, t2.b
FROM t1, t2
WHERE t1.a = t2.a
AND  t1.c = y
  AND t1.f IN (n, o, p)
  AND t2.d = v
  AND t2.e BETWEEN k AND h
UNION
SELECT t1.a, t2.b
FROM t1, t2
AND t1.c = z
  AND t1.f IN (p, q)
  AND t2.d = w
  AND t2.e BETWEEN k AND h
```

©Copyright 2004 OSDL, All rights reserved.

Query #19 Takes Too Long To Execute

```
•select sum(l_extendedprice* (1 - l_discount)) as revenue
 from   lineitem, part
 where  (
          p_partkey = l_partkey
          and p_brand = ':1'
          and p_container in ('SM CASE', 'SM BOX', 'SM PACK', 'SM PKG')
          and l_quantity >= :4 and l_quantity <= :4+10
          and p_size between 1 and 5
          and l_shipmode in ('AIR', 'AIR REG')
          and l_shipinstruct = 'DELIVER IN PERSON'
        )
        or
        (
          p_partkey = l_partkey
          and p_brand = ':2'
          and p_container in ('MED BAG', 'MED BOX', 'MED PKG', 'MED PACK')
          and l_quantity >= :5 and l_quantity <= :5+10
          and p_size between 1 and 10
          and l_shipmode in ('AIR', 'AIR REG')
          and l_shipinstruct = 'DELIVER IN PERSON'
        )
        or
        (
          p_partkey = l_partkey
          and p_brand = ':3'
          and p_container in ('LG CASE', 'LG BOX', 'LG PACK', 'LG PKG')
          and l_quantity >= :6 and l_quantity <= :6+10
          and p_size between 1 and 15
          and l_shipmode in ('AIR', 'AIR REG')
          and l_shipinstruct = 'DELIVER IN PERSON'
        );
```

©Copyright 2004 OSDL, All rights reserved.

Rewritten Query #19

```
•select sum(l_extendedprice* (1 - l_discount)) as revenue
 from   lineitem, part
 where
        p_partkey = l_partkey
        and l_shipmode in ('AIR', 'AIR REG')
        and l_shipinstruct = 'DELIVER IN PERSON'
        and
        (
          (
             p_brand = ':1'
             and p_container in ('SM CASE', 'SM BOX', 'SM PACK', 'SM PKG')
             and l_quantity >= :4 and l_quantity <= :4+10
             and p_size between 1 and 5
          )
          or
          (
             p_brand = ':2'
             and p_container in ('MED BAG', 'MED BOX', 'MED PKG', 'MED PACK')
             and l_quantity >= :5 and l_quantity <= :5+10
             and p_size between 1 and 10
           )
           or
           (
             p_brand = ':3'
             and p_container in ('LG CASE', 'LG BOX', 'LG PACK', 'LG PKG')
             and l_quantity >= :6 and l_quantity <= :6+10
             and p_size between 1 and 15
          )
        );
```

©Copyright 2004 OSDL, All rights reserved.

Query #19 Results

- Using 7.5
  - Modified Q19: 96 seconds in the Power Test
  - Original Q19: 54 seconds in the Power Test

©Copyright 2004 OSDL, All rights reserved.

Additional Benefits from Testing Q19

- To make results of DBT-3 more useful for developers, modified the test to run EXPLAIN ANALYZE instead of returning results for each query.
- EXPLAIN ANALYZE data lead to more optimizer discussions for two other queries

©Copyright 2004 OSDL, All rights reserved.

Q21

```
select s_name, count(*) as numwait
from   supplier, lineitem l1, orders, nation
where  s_suppkey = l1.l_suppkey
  and o_orderkey = l1.l_orderkey
  and o_orderstatus = 'F'
  and l1.l_receiptdate > l1.l_commitdate
  and exists ( select *
               from   lineitem l2
               where l2.l_orderkey = l1.l_orderkey
                 and l2.l_suppkey <> l1.l_suppkey )
  and not exists ( select *
                   from   lineitem l3
                   where l3.l_orderkey = l1.l_orderkey
                     and l3.l_suppkey <> l1.l_suppkey
                     and l3.l_receiptdate > l3.l_commitdate )
  and s_nationkey = n_nationkey
  and n_name = 'MOROCCO'
group by s_name
order by numwait desc, s_name
LIMIT 100;
```

©Copyright 2004 OSDL, All rights reserved.

Analysis

- "It seems that the correct way to plan this query would require postponing evaluation of the EXISTS clauses. If those were further up the tree, the planner would have chosen a merge or hash join at this step, which would probably take a tenth as much time. The cost to run the EXISTS clauses themselves wouldn't change; they'd not be executed any more frequently in this case." - Tom Lane

©Copyright 2004 OSDL, All rights reserved.

Q4

- ```
  where o_orderdate >= date '1995-04-01'
    and o_orderdate < date '1995-04-01' + interval '3 month'
  ```
- "As of CVS tip the issue could be eliminated by introducing cross-data-type comparison operators between types date and timestamp without time zone, and then making these be members of the date index opclass. I'm strongly tempted to do so ..." - Tom Lane

  ....and he did!

©Copyright 2004 OSDL, All rights reserved.
  ```

Kit info

- Sourceforge:

  http://sourceforge.net/projects/osdldbt
- You can use the kit too!
  - Run as is on your version of Linux/hardware to measure performance
  - Customize the queries/stored procedures
  - Changing the schema would be harder (data generation issues).
- DBT tests run on our automated test platform -
  - You can look at
    - Results of Linux kernel testing
    - Results of new PostreSQL releases
    - http://osdl.org/stp

©Copyright 2004 OSDL, All rights reserved.