

Solaris pNFS Client Works In Progress

Bill Baker

Rich Brown

Sun Microsystems



Overview

- Client side control of layouts
- Layout flows and strategies
- Where's my pNFS?
- Dtrace providers

User Layout Control

- Allows the user to specify desired layouts
- Attributes that determine which policy to use:
 - > path
 - > uid, gid
 - > file extension (e.g, .jpeg, .mpeg)
 - > time, date
- Server may override or simply ignore :-)
- No application changes required, transparent

Example

- `nfsadm policy-get -c all`

| NAME | STRIPE_COUNT | UNIT_SIZE | NPOOLS | RULE |
|---------|--------------|-----------|--------|------|
| default | 1 | 16k | - | - |

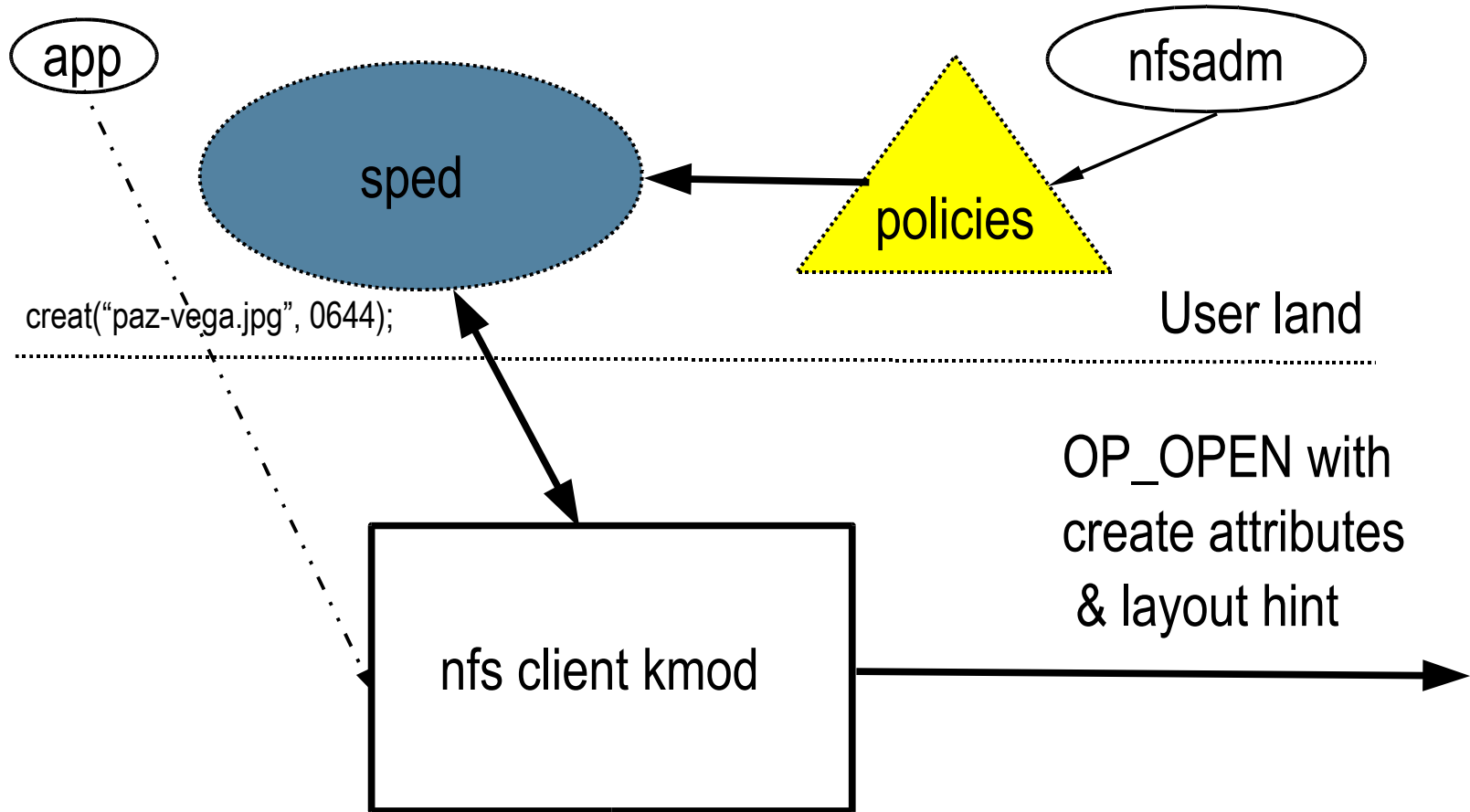
- `nfsadm policy-create -c -o stripe-count=2,unit-size=32k,rule=path=/pnfs/cool-stuffs 2way`

- `nfsadm policy-get -c all`

| NAME | STRIPE_COUNT | UNIT_SIZE | NPOOLS | RULE |
|---------|--------------|-----------|--------|------------------------|
| default | 1 | 16k | - | - |
| 2way | 2 | 32k | - | path=/pnfs/cool-stuffs |

- `cp /home/rmesta/paz-vega.jpg /pnfs/cool-stuffs`

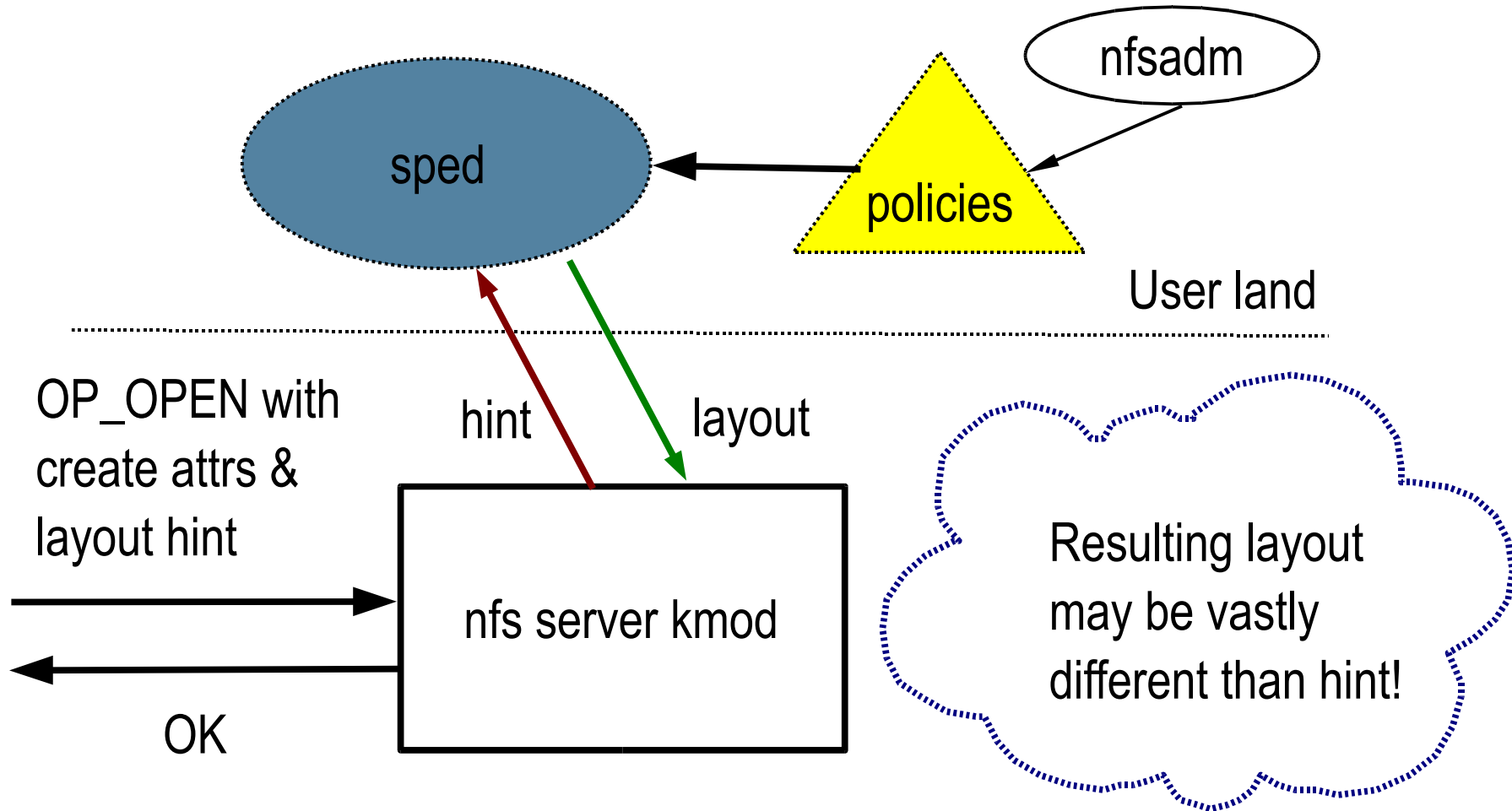
Layout Flow – pNFS client



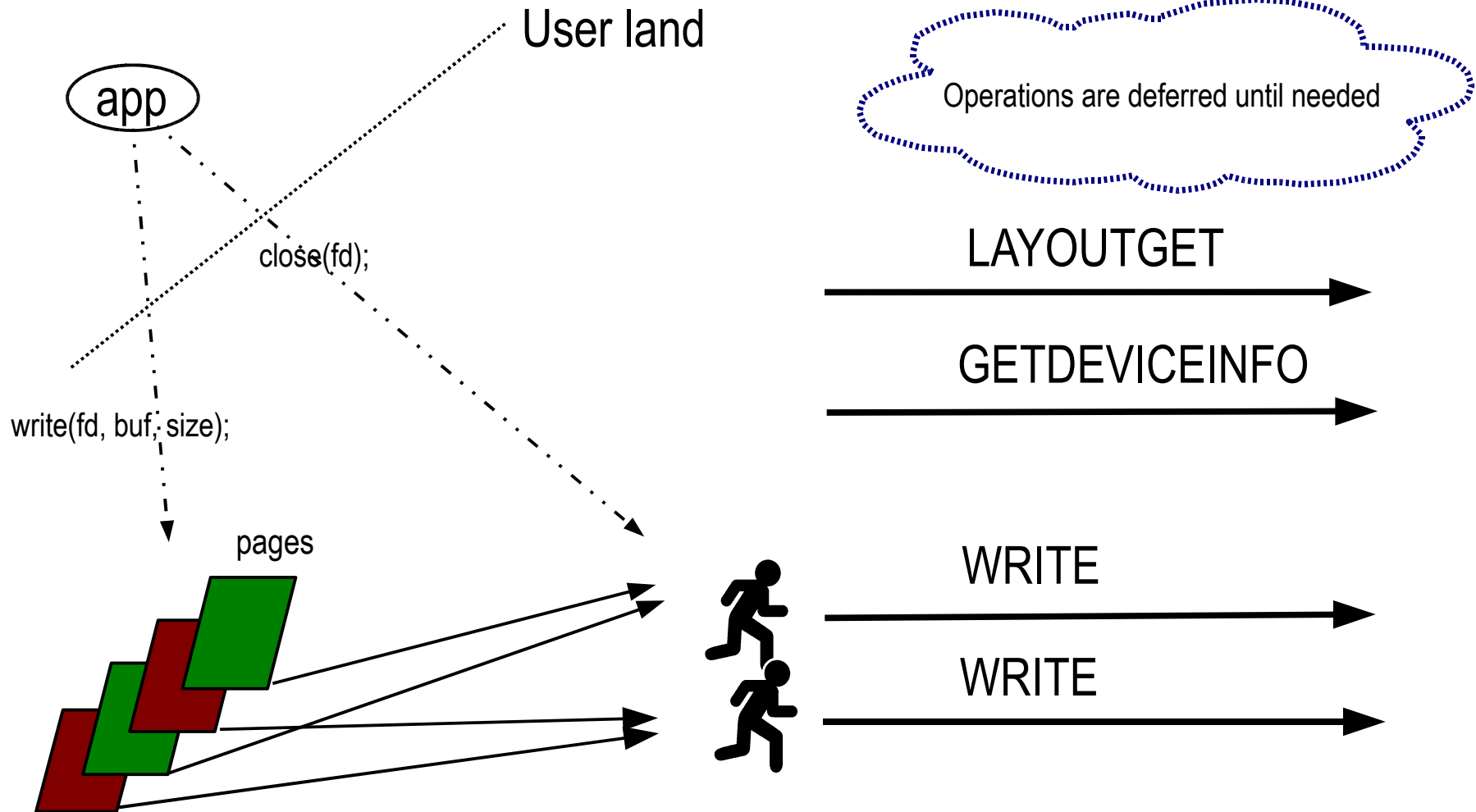
Layout hint, snoop -v

```
NFS: Op = 18 (OPEN)
NFS: Claim Type = 0 (NULL)
NFS: paz-vega.jpg
NFS: Open Type = CREATE
NFS: Method = GUARDED
NFS: 0x10  SIZE
NFS: 0x00
NFS: 0x00
NFS: 0x00
NFS: 0x02  MODE
NFS: 0x00
NFS: 0x00
NFS: 0x00
NFS: Size = 0
NFS: Mode = 0644
NFS: Client specified hint for file layout: LAYOUT4_NFSV4_1_FILES
NFS:  care = 0xc  CARE_STRIPE_UNIT_SIZE  CARE_STRIPE_COUNT
NFS:  util = 0x8000
NFS:  stripe_count = 2
```

Layout Flow – pNFS MDS



Layout Flow – pNFS client




Am I getting pNFS?

```
$ nfsstat -c -v 41
```


```
Version 41: (145 calls)
```

```
null          compound
0 0%          145 100%
```

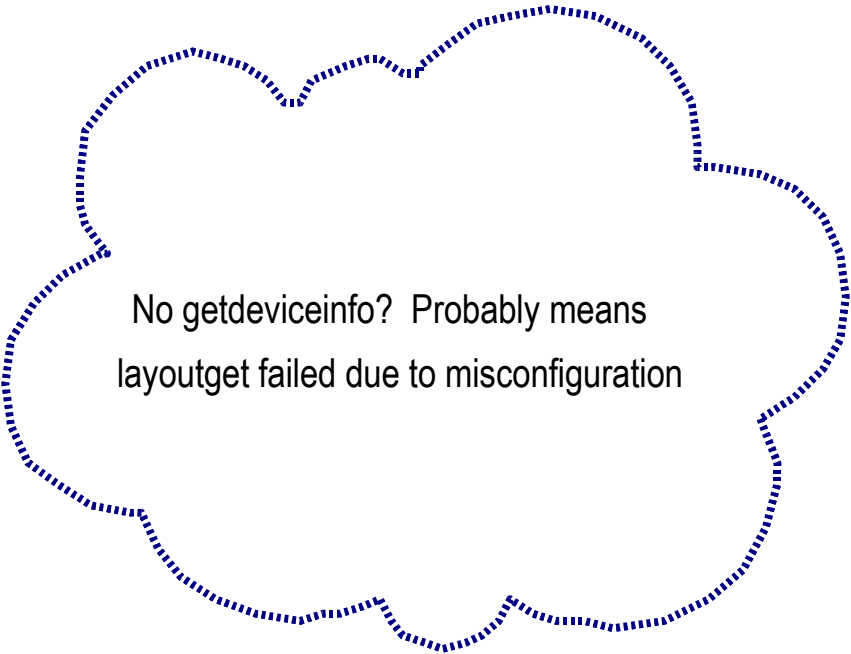
```
<...>
```

```
bind_conn_to_session exchange_id   create_session
0 0%                3 2%  3 3%
```

```
destroy_session  free_stateid   get_dir_delegation
0 0%             0 0%           0 0%
```

```
getdeviceinfo   getdevicelist  layoutcommit
1 0%  0 0%           3 0%
```

```
layoutget       layoutreturn   secinfo_no_name
2 0%            1 0%           0 0%
```



No getdeviceinfo? Probably means layoutget failed due to misconfiguration

What's my layout?

- `nfsstat -l /pnfs/cool-stuffs/paz-vega.jpg`

Number of layouts: 1

Proxy I/O count: 0

DS I/O count: 2

Layout [0]:

Layout creation timestamp: Wed Feb 18 19:09:29:353122 2009

Layout [0]:, iomode: LAYOUTIOMODE_RW

offset: 0, length: EOF

num stripes: 2, stripe unit: 32768

Stripe [0]:

tcp:salma:172.20.27.126:4920 OK

Stripe [1]:

tcp:jlo:172.20.48.137:44722 OK

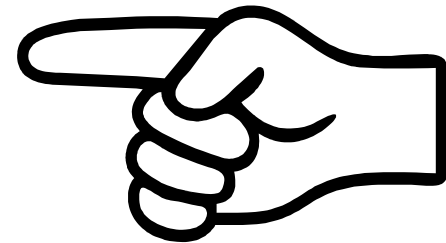


What if the client doesn't have a layout?

LAYOUTGET Strategies

When Should the client do LAYOUTGET?

- As part of the open compound?
- Asynchronously, initiated during open?
- Lazily, only when needed?
- Differing strategies for read vs. write?
- Differing strategies for create vs. existing?



We're willing to do the bare minimum

Dtrace Providers for nfs

Why a provider?

Abstraction from code structure

Argument aggregation

Better than snoop



snoop krb5p? Good
luck

<http://wikis.sun.com/display/DTrace/nfsv4+Provider>

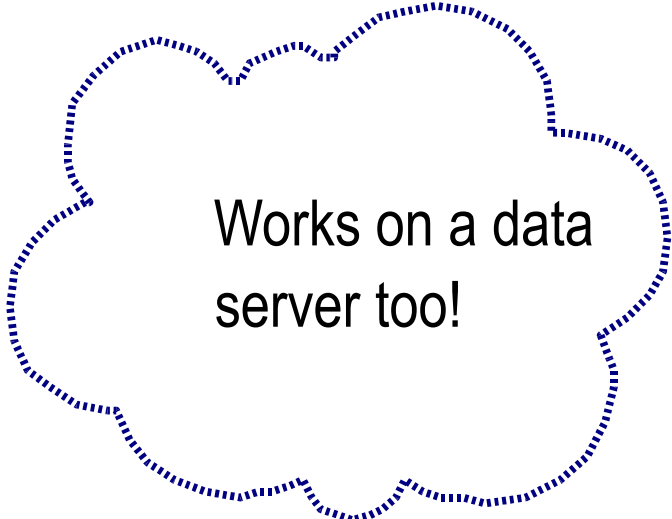
<http://wikis.sun.com/display/DTrace/nfsv3+Provider>

Dtrace Provider for NFSv4 server


```
#!/usr/sbin/dtrace -s

nfsv4:::op-read-start
{
    self->t0 = timestamp;
}

nfsv4:::op-read-done
/ args[2]->status == 0 /
{
    rr = (READ4res *) args[2];
    printf("read %d bytes in %d microseconds\n",
        rr->data_len, (timestamp-self->t0)/1000);
}
```



Works on a data
server too!



Available in
OpenSolaris now!

Dtrace Provider for NFSv4 client

```
#!/usr/sbin/dtrace -s

nfsv4-client:::compound-op-start
{
    self->t0 = timestamp;
}

nfsv4-client:::compound-op-done
/ self->t0 /
{
    c4res = (COMPOUND4res *) args[2];
    printf("compound %s took %d microseconds\n",
        c4res->tag, (timestamp - (self->t0))/1000);
}
```

Questions?

nfsv41-discuss@opensolaris.com

Badges?

“We don't need no stinking badges!”

Solaris pNFS Client Works In Progress

Bill Baker

Rich Brown

Sun Microsystems

