13th ANNUAL WORKSHOP 2017

MANAGING NODE CONFIGURATION WITH 1000S OF NODES

Ira Weiny

Intel Corp

2017

PROBLEM

- **Clusters are built around individual “servers”**
- **Linux configuration is often designed around a single desktop or single server**
- **Single server configuration can present problems with running clusters at scale**
- **Various rdma tools now exist to aid in managing nodes**

IBACM

IBACM

What is it

- **From the man page:**

“The IB ACM implements and provides a framework for name, address, and route (path) resolution services over InfiniBand….”

… and Omni-Path Architecture

“The IB ACM package is comprised of three components: the ibacm core service, the default provider ibacmp shared library, and a test/configuration utility - ib_acme. All three are userspace components and are available for Linux. Additional details are given below.”

IBACM

What is it, and why use it?

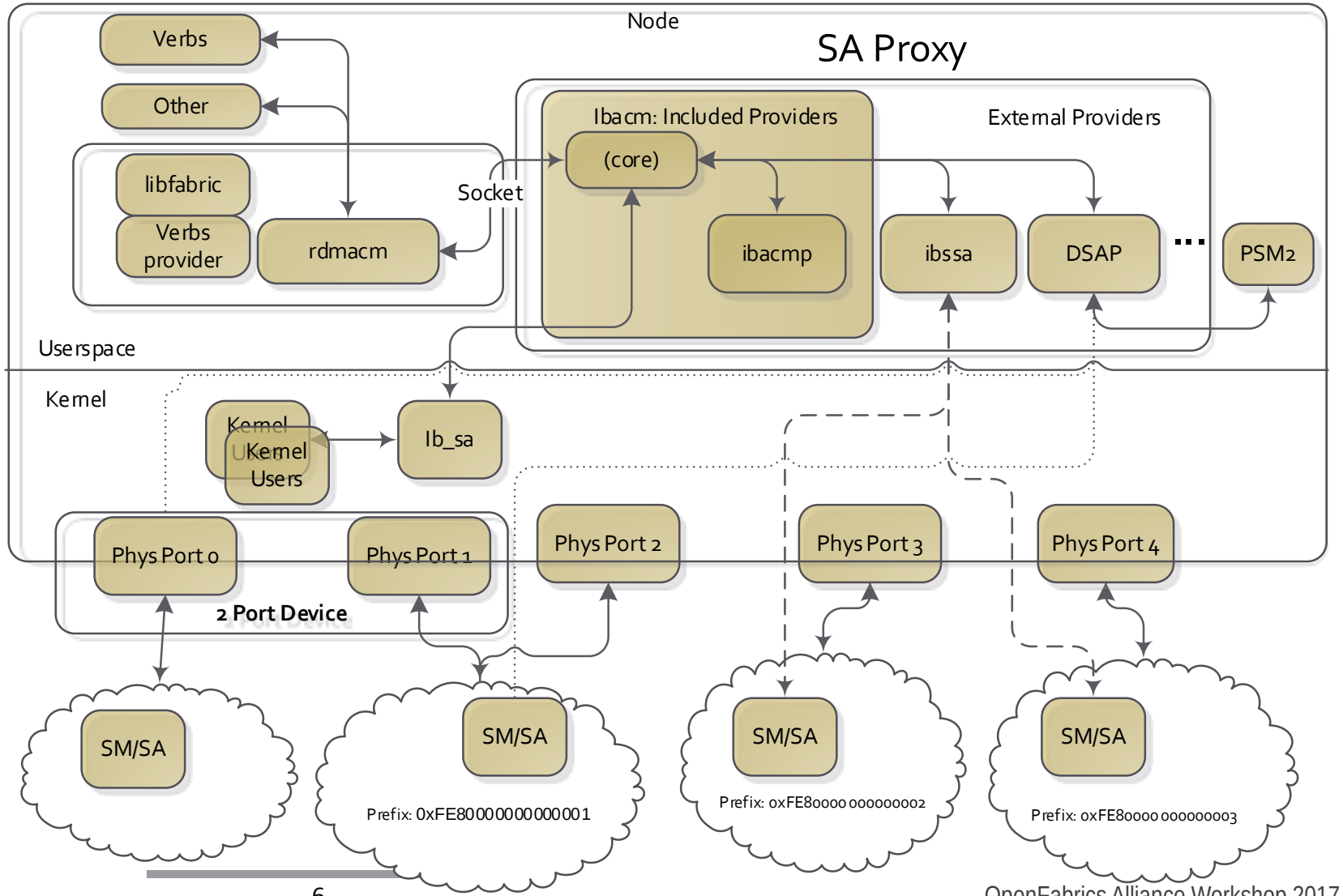
- **Ibacm and its companion librdmacm are essential to generic fabric operation**

- **Doesn’t ibacm have scalability issues?**

NO! not when used properly!

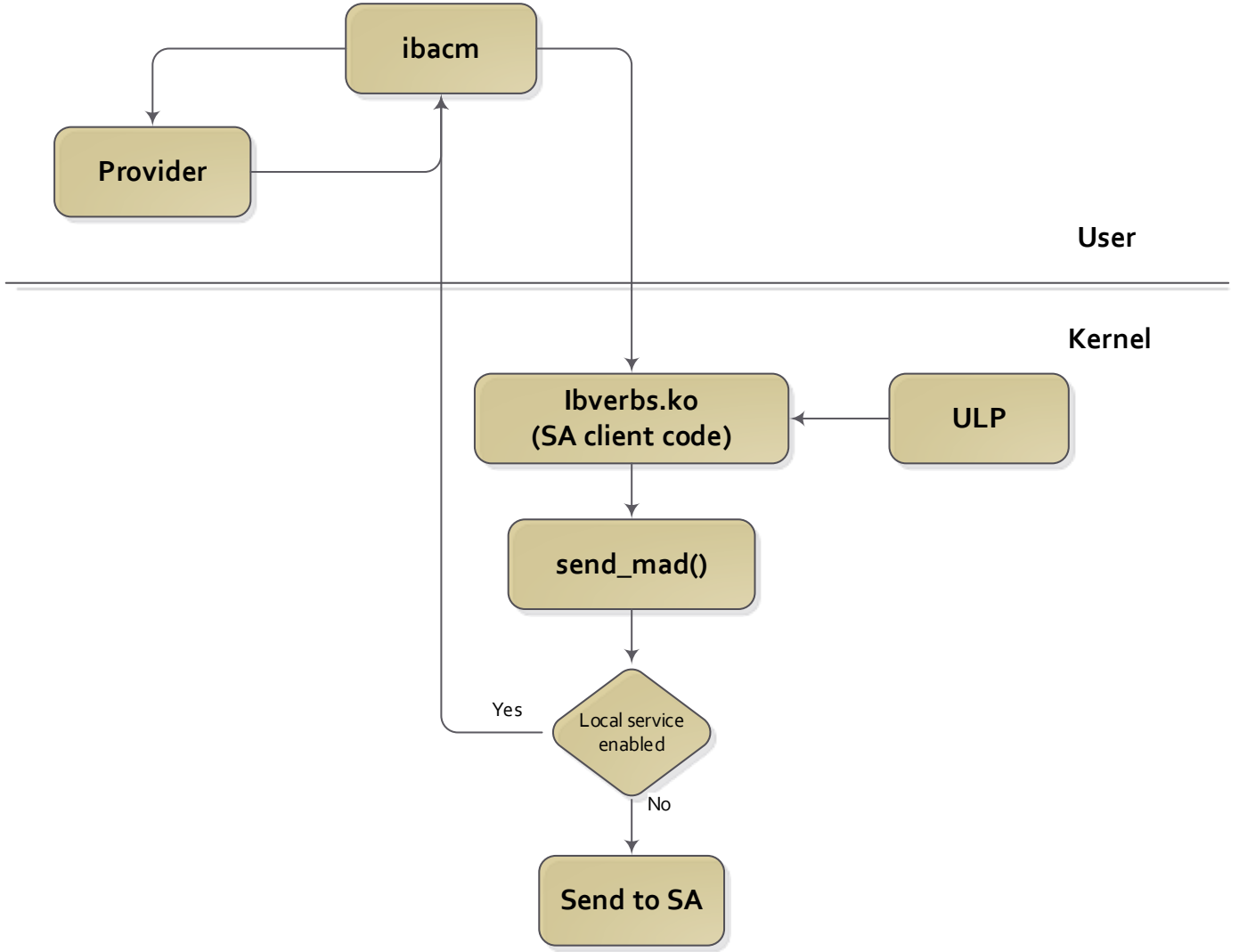
IBACM

- **Local SA daemon**
- **Now with plugins!**



IBACM
Kernel Support

- **Local SA daemon**
- **Now with kernel support!**
- **Currently only supports IB Path Records**
- **Planned support for**
  - OPA Path Records
  - Multicast Records



IBACM
Installation

- **How do I get it?**
  - Your friendly neighborhood distro
    - RHEL, SLES, Ubuntu
  - Rdma-core
    - https://github.com/linux-rdma/rdma-core
- **What version do I need**
  - V1.1.0 or greater
    - NOTE: the rdma-core version of the package jumps to match the rdma-core version
  - Kernel support was in added in 4.3

IBACM

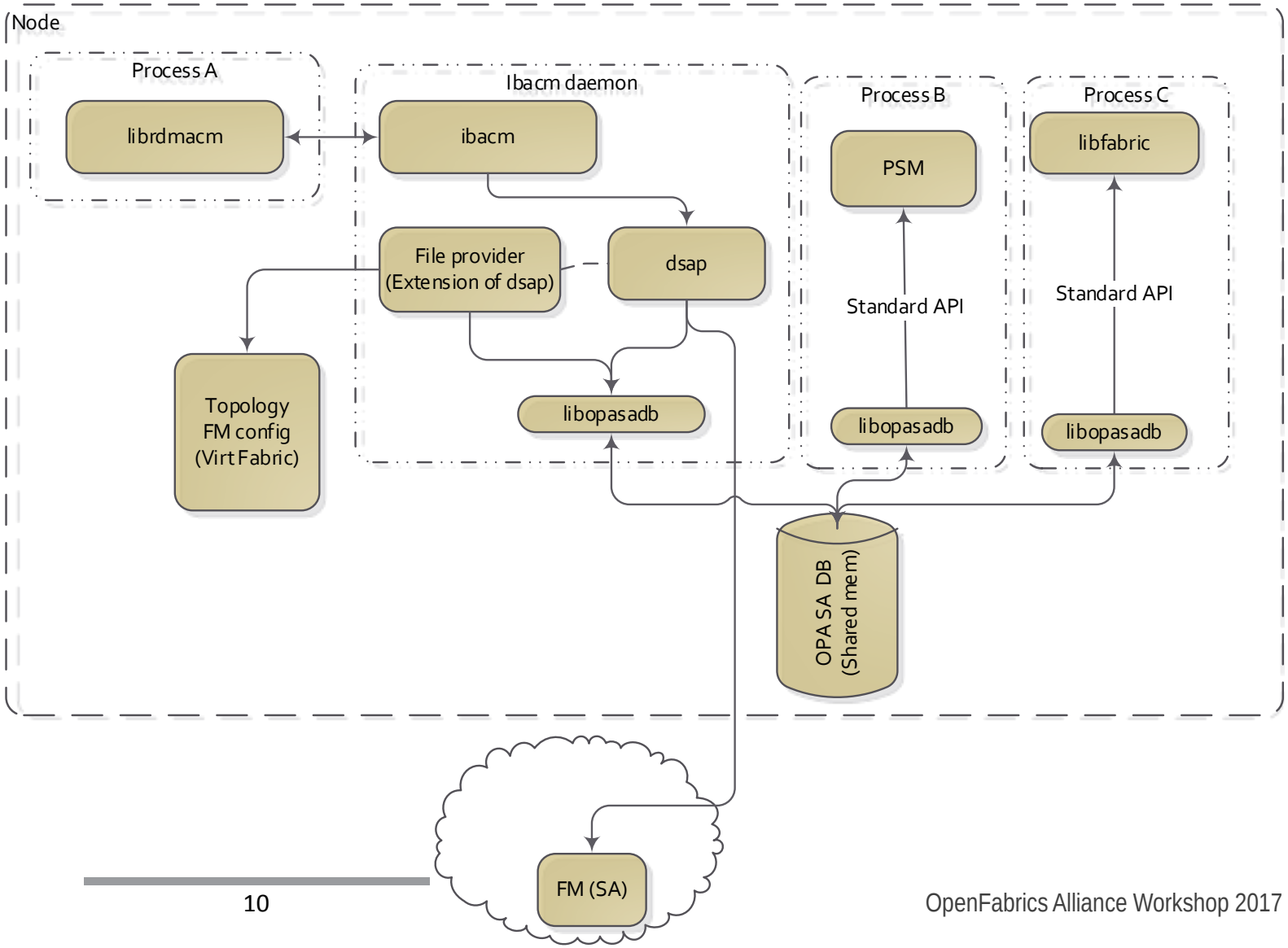
Plugins

- **IB**
  - Acmp – Default provider included in ibacm (rdma-core) package
  - Ibssa – external provider for opensm
- **OPA**
  - "DSAP" – Distributed SA Provider
    - Part of the opa-ff package
      - https://github.com/01org/opa-ff
    - Rpm name Opa-address-resolution

IBACM

DSAP provider (plugin)

- **Provides standard Path Record lookups to librdmacm users**
  - Typically verbs users
- **Also provides shared memory access to it's internal database through "libopasadb"**



IBACM

- **Ensure IPoIB is configured and running on the ports you want to control**

```
$ ifconfig ib0
ib0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 65520
        inet 10.228.217.70  netmask 255.255.252.0  broadcast 10.228.219.255
        inet6 fe80::211:7501:165:abf8  prefixlen 64  scopeid 0x20<link>
Infiniband hardware address can be incorrect! Please read BUGS section in ifconfig(8).
        infiniband 80:00:00:0E:FE:80:00:00:00:00:00:00:00:00:00:00:00:00:00:00  txqueuelen 256
(InfiniBand)
        RX packets 8  bytes 504 (504.0 B)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 19  bytes 3122 (3.0 KiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

IBACM

- **Configure your providers**

<edit /etc/rdma/ibacm_opts.cfg to assign providers to specific subnets>

```
# provider:
# Specifies the provider to assign to each subnet
# ACM providers may override the address and route resolution
# protocols with provider specific protocols.
# provider name (prefix | default)
# Example:
provider dsap 0xFE80000000000000
provider ibacmp default
```

- **If the config file does not yet exist use ib_acme to generate one**

```
$ ib_acme -O /etc/rdma/ibacm_opts.cfg
```

IBACM

- **Ensure ibacm is configured to run on startup and restart the daemon**

$ systemctl enable ibacm

$ systemctl restart ibacm

IBACM

- **Test a query**

$ ib_acme -f lid -s 2 -d 1

```
Service: localhost
Destination: 1
Source: 2
Path information
  dgid: fe80::11:7501:165:ae8b
  sgid: fe80::11:7501:165:abf8
  dlid: 1
  slid: 2
  flow label: 0x0
  hop limit: 0
  tclass: 0
  reversible: 1
  pkey: 0x8001
  sl: 0
  mtu: 7
  rate: 16
  packet lifetime: 15

return status 0x0
```

IBACM

Verbs Access

- **Use librdmacm for connection management**
- **Ibacm is automatically contacted from the library and cached path records are utilized!!!**

- **Configure PSM2 to use Path Record queries**

<add the following to mpi_run scripts or otherwise add the environment variable setting>

export PSM2_PATH_REC=OPP

IBACM

- **What happens if ibacm and/or dsap is not running/configured?**

phcppriv12.ph.intel.com.75727ips_opp_init: PSM path record queries using OFED Plus Plus (libopasadb.so.1.0.0) from /lib64/libopasadb.so.1.0.0

2017-03-12 23:46:23| ERROR: Unable to open shared memory table.

phcppriv12.ph.intel.com.75727Unable to initialize OFED Plus library successfully.
(err=51)

phcppriv12.ph.intel.com.75727mpi_stress: OPP: Unable to obtain OPP context. Disabling OPP interface for path record queries.

phcppriv12.ph.intel.com.75727mpi_stress: Make sure SM is running...

phcppriv12.ph.intel.com.75727mpi_stress: Make sure service ibacm is running...

phcppriv12.ph.intel.com.75727mpi_stress: to start ibacm: service ibacm start

phcppriv12.ph.intel.com.75727mpi_stress: or enable it at boot time: opaconfig -E ibacm

IBACM
Other Details

- **Log files**
  - /var/log/ibacm.log
- **Config files**
  - /etc/rdma/ibacm_opts.cfg
  - /etc/rdma/ibacm_addr.cfg

IBACM

Further work

- **Kernel support**
  - Add support for OPA Path Records
  - Add support for Multicast Records
  - Configure start up to ensure ibacm can run prior to kernel ULPs
- **ibacm/providers**
  - Multicast group operations (join/leave)
    - Pre-joined groups
  - Multicast Records
  - Preconfigured topologies
  - Remove IPoIB dependency

RDMA-NDD

RDMA-NDD

Keeping nodes named so you don't have to

- **Setting node descriptions has a long history**
- **First there was the scripts to be run at start up…**
  - set_nodedesc.sh -- 2007
  - RHEL udev rule – RHEL 6 time?
  - Truescale IFS – hard coded in driver
  - rdma-ndd – 2014
- **rdma-ndd was built to be flexible and robust**
  - Watches for hostname changes
  - Watches for device adds
  - Responds to these events by setting the node description

RDMA-NDD

Where do I get it

- **Friendly neighborhood distro!**
  - At least some distros have integrated it with their udev rules
  - NOTE: It is separated to it's own rpm: rdma-ndd-*.rpm
- **Upstream**
  - infiniband-diags > v1.6.5
    - Soon to be removed
  - rdma-core-12 or greater

RDMA-NDD

Configuration

- **Configuration differs depending on the version**
  - Infiniband-diags version uses the ibdiag.conf file
  - Rdma-core version is configured within systemd
- **The format can take 2 wild cards**
  - %h => replace with the hostname
  - %d => replace with the device name
- **Default is “%h %d”**

SUMMARY

- **Two daemons are now standard as part of the rdma-core**

ibacm

rdma-ndd

- **These daemons make it easier for clusters of nodes to be managed together**

13th ANNUAL WORKSHOP 2017

THANK YOU

Ira Weiny

Intel Corp

DISCLAIMERS

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

Forecasts: Any forecasts of requirements for goods and services are provided for discussion purposes only. Intel will have no liability to make any purchase pursuant to forecasts. Any cost or expense you incur to respond to requests for information or in reliance on any forecast will be at your own risk and expense.

Business Forecast: Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

Intel, the Intel logo, are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others