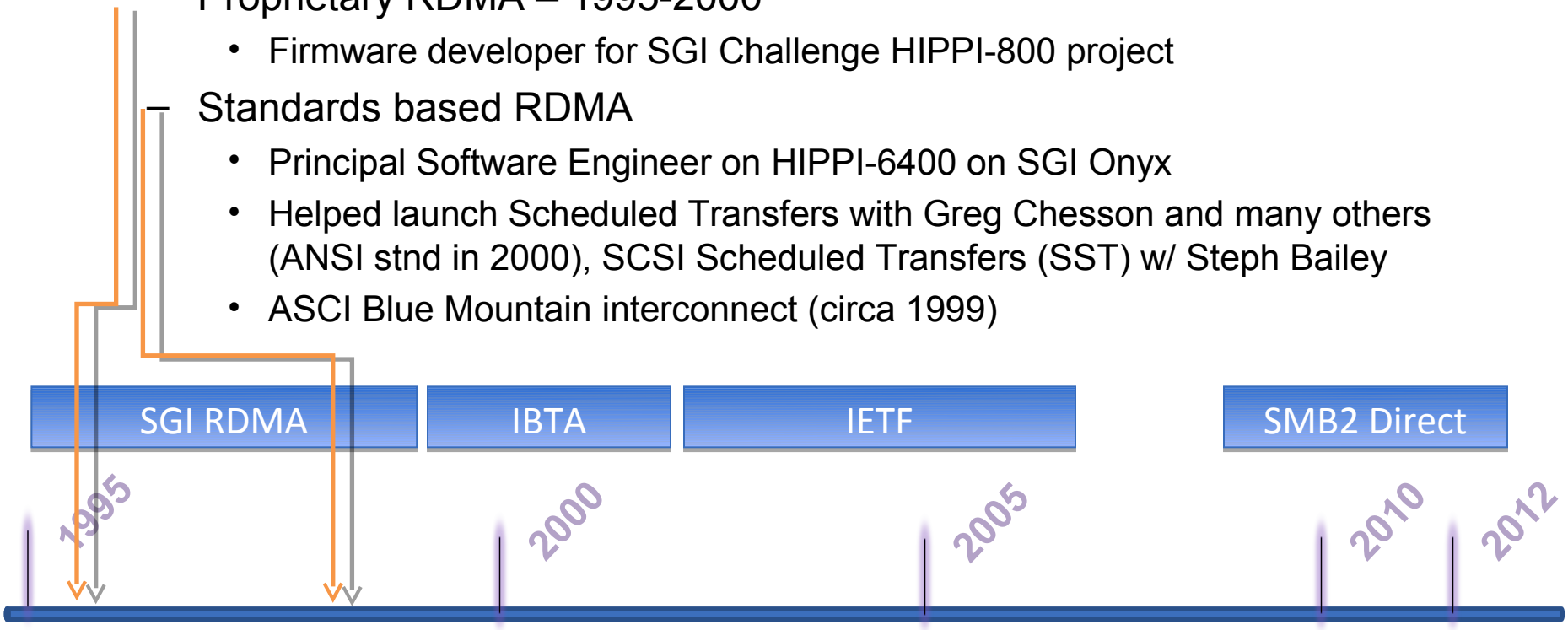
# What a Long Strange Trip It's Been:

# Moving RDMA into Broad Data Center Deployments

Author: Jim Pinkerton, Partner Architect, Microsoft
Date: 3/25/2012

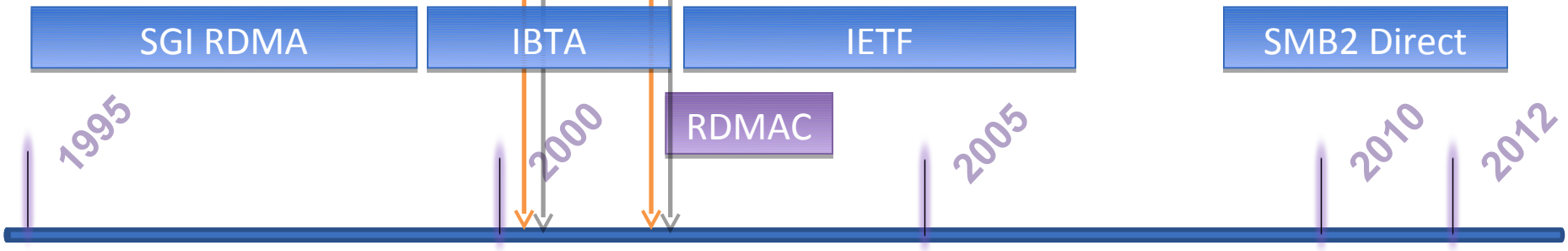
# What a Long Strange Trip

- Who am I? - an old fart
- SGI
  - Proprietary RDMA – 1995-2000
    - Firmware developer for SGI Challenge HIPPI-800 project
  - Standards based RDMA
    - Principal Software Engineer on HIPPI-6400 on SGI Onyx
    - Helped launch Scheduled Transfers with Greg Chesson and many others (ANSI stnd in 2000), SCSI Scheduled Transfers (SST) w/ Steph Bailey
    - ASCI Blue Mountain interconnect (circa 1999)



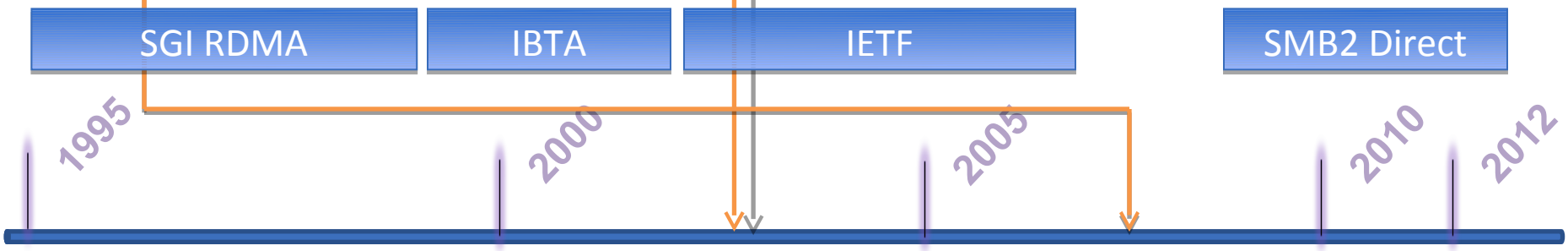
# What a Long Strange Trip

- Microsoft: 2000 until now
  - IBTA: Co-Chair Software Working Group, on board of IBTA
    - Co-Author of Sockets Direct Protocol (SDP) 2002
    - Herded cats on IBTA Verbs 1.0
  - RDMA Consortium: helped found it. Co-chair working group
    - Co-Author Verbs port to iWARP, SDP port to iWARP, DDP
    - RDMA Consortium (RDMAC) formed **5/02**
    - RDMAC RDMAP, DDP, MPA completed **10/02**
    - RDMAC Verbs completed **4/03**
    - iSCSI Extensions for RDMA (iSER) **7/03**, Sockets Direct Protocol (SDP) **10/03**



# What a Long Strange Trip

- Microsoft (still there…)
  - IETF RDDP Work Group Chartered 2003
    - Co-author of RDMA security draft **10/07**
    - RDMAC core specs become RFCs, with minor changes **10/07**
      - (except SDP and Verbs)
    - iSER becomes RFC – **10/07**
    - IPS Working Group Disbanded – **11/07**
    - STORM WG chartered - **2009**
  - Open Fabrics
    - Helped charter Windows OFED (Windows Working Group) with Windows HPC team
  - Transitioned to learn storage for real…



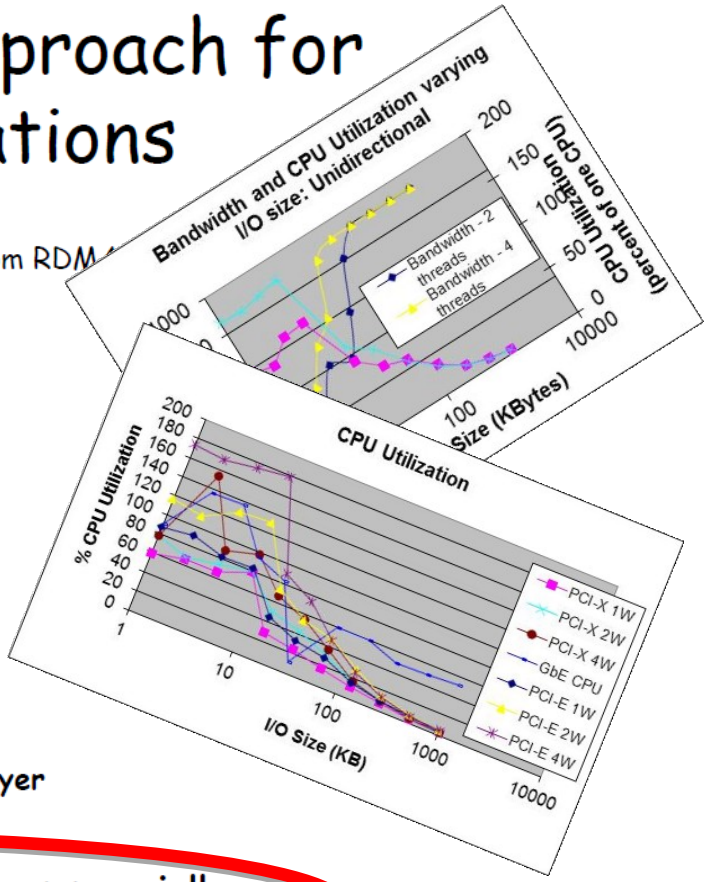
# I thought I knew what I was doing…

OPENFABRICS ALLIANCE

## RDMA is a proven approach for Today's Applications

5/29/2002
RDMA Consortium
Tutorial

- Transactional Database
  - SQL, DB2, Oracle all get best performance from RDM
  - SDP – Sockets Direct Protocol – is also used
- File Oriented Storage
  - DAFS, BDS (SGI), NFS (research by Sun)
- Block Oriented Storage
  - ANSI SRP, ANSI SST
- High Performance Computing
  - Various MPI libraries based on VIA
  - ASCI Blue Mountain (SGI)
- Backup
  - Backup system from SGI




- Summary
  - All met their targeted performance goals
  - All are currently limited to the Data Link Layer
  - All the above are commercial products

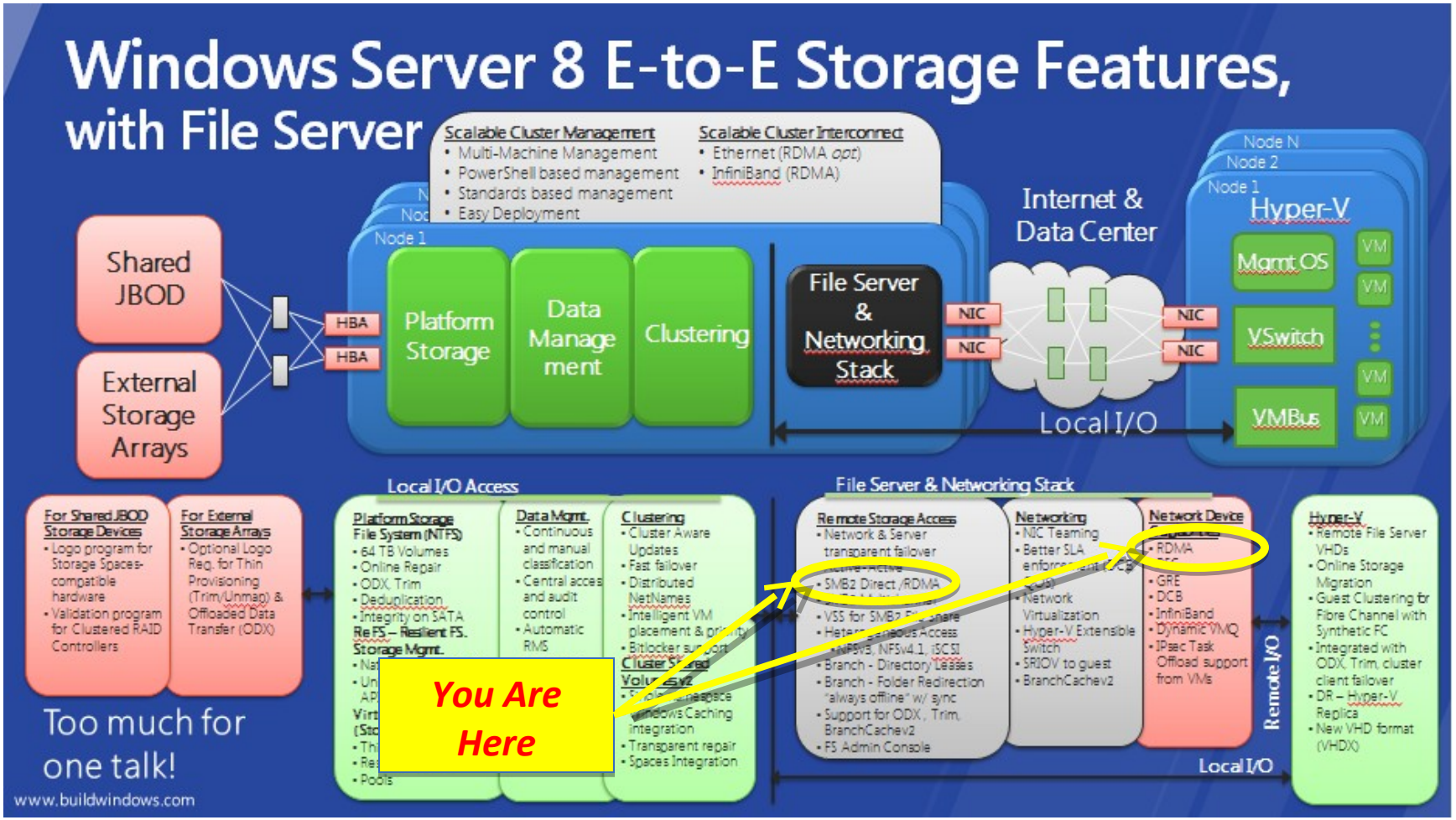
- Almost none of the above are/were commercially successful

# My Goals – For Far Too Long…

- Low overhead network protocol on both high speed phy and volume phy
  - Overhead can be CPU, latency, …
- Make it useful to the masses – not just clusters
- Find an app

- But I missed the system…

# The System




From www.buildwindows.com talk SAC-446

# Requirements for RDMA Broad Data Center Deployment (1/3)

- Goal is to enable data center scale, one fabric
  - Not contained to clusters within the data center
  - Not with two networks
- Must interact well with existing data center management infrastructure
  - Encapsulate IP traffic
  - Support mapping of ethernet protocols
    - IGMP, ARP, Broadcast, Multicast, VLAN, NVGRE
  - Integrate with system statistics, performance counters
  - Compatible with firewall rules
  - Supports remote boot environment

# Requirements for RDMA broad Data Center Deployment (2/3)

- Network must be able to be highly fault tolerant
  - Preferably active-active with transparent failover
- Must be extremely easy to deploy
  - On by default
- Security is critical
  - No buffers can be handed to an application that are still able to be modified from the network

# Requirements for RDMA broad Data Center Deployment (3/3)

- Must have an end-to-end scenario/application
  - I chose file-based storage
  - Storage is “different” – Storage latencies are not HPC latencies
    - BW focus is large I/Os (I chose 512 KB)
    - IOPs focus is small I/Os (I chose 8 KB)
    - Latency focus is log file (I chose 1 KB writes)

# SMB2.2 Requirements of SMB Direct (SMB over RDMA)

- Goal for SMB2.2:
  - Remote file storage similar to local in functionality, performance, reliability, availability
  - RDMA was critical to achieve performance goals
- SMB2.2:
  - Integrated with application snapshots
  - Transparent file server node & network failover
    - Bounded time (<<25 sec), under full load (20K clients)
  - **CPU overhead the same as local attached storage**
  - **Applicable for cluster storage interconnect and file serving to server applications**

# SMB2.2 Requirements of SMB Direct (SMB over RDMA)

- Scalable – multiple interfaces, multiple connections per interface
  - Active-active with transparent failover of interfaces
  - Multiple connections per interface
- Virtualization and live migration
  - Map local storage within a VM to remote file storage
  - On moving a VM, be able to transparently reconnect even if the only interface is non-RDMA
  - On moving a VM, be able to transparently upgrade if RDMA capable NIC appears

# On Bring Up…

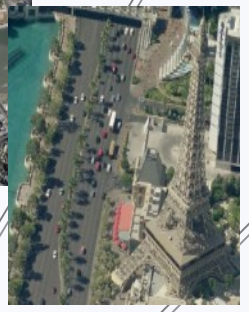
- Large I/Os hit the goals immediately after the system started working
  - Serious system bottlenecks (pre Sandy Bridge)
- Small I/Os we spent a year tuning
  - 8 KB are zero-copy (RDMA)
  - 1 KB uses Send/Rcv (one copy)
- Latency not as big a focus because of non-volatile storage focus (i.e. storage is slow)
- Supporting 3 RDMA fabrics requires swatting a lot of bugs on a per-vendor basis…

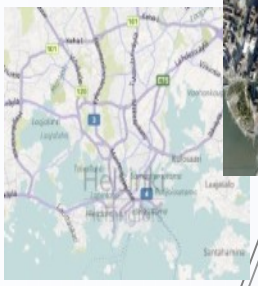
# SMB Direct Beta Deployment: Bing Maps

**AERIAL**

Global Ortho

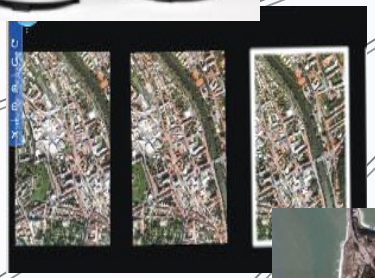
Oblique



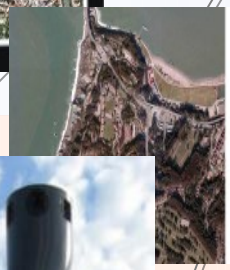


Vectors

UltraCam Aerial



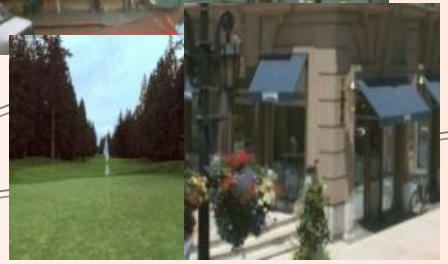
UltraMap



UltraCamM
UltraCamP

**Graz, Austria**
UltraCam development
Commercial software
Computer vision research

BlockView



Streetside

**Boulder, Colorado**
Bing Maps data center
Software development
Acquisition management

**TERRESTRIAL**

# SMB Direct Beta Deployments Bing Maps

- Compute farm with massive data ingest
  - Extreme focus on efficiency & cost
  - Phased deployment of Windows Server “8”
- The Bing Maps data center
  - Multiple containers
  - ~2000 nodes, ~20K cores, **55 PB of data**
  - Growing quickly
  - Transitioned from 1 gigabit network to 100% QDR InfiniBand
  - Every node is a file server and a compute engine
  - Initially IP over IB
  - SMB Direct in beta (just 44 nodes today)
    - ~17 gbits/sec throughput in production today (disk to network to memory), 512 KB IO, one outstanding IO

# Other Deployments

- Microsoft internal VM farm (iWARP)
  - Runs automated tests for Windows
  - Everything is virtualized
  - Just starting to deploy
- Several others that aren't public yet
  - ROCE, iWARP, InfiniBand

# Open Protocol Documentation

- All SMB2 protocols published, including RDMA
  - Bing for “ms-smbd protocol” ☺
- Protocol Family
  - Existing docs being updated
    - MS-SMB2 – SMB2
    - MS-DFSC – DFS Namespaces
    - MS-FSCC – (File System Control Codes)
    - MS-FSA – File System Algorithms
  - New Protocol Documents
    - **MS-SMBD – SMB2 Direct (RDMA)**
    - MS-FSRVP – Remote VSS Protocol
    - MS-SWN – SMB Witness Protocol

**See Tom Talpey's talk on SMB Direct on Tuesday at 8:00am.**

# Where do I want to go?

- RDMAv2
  - iWARP overhead on RDMA Reads is depressing
  - Better multi-tenancy
  - I want a reliable RDMA write
- While RDMA has substantially lowered the cost of remote I/O, with the goal that “remote is the same as local”, local I/O is terrible
  - Disk latency: ~1-10 msec
  - Volume Flash latency: ~200 usec – 1 msec (95$^{th}$ percentile)
- **NVME to the rescue?**

# What have I learned?

I can be persistent (or stubborn)

Application have to go native RDMA

Technology solutions are tough to sell