



CANARIE's Wavelength Disk Drive Project : Experiences from Phase I and Outlook for the Future

<http://www.canet3.net/wdd/>

CANARIE Inc. : Bill St. Arnaud, René Hatem

Can-Sol Computer Corporation : Rick Ingram

Carleton University : Doron Nussbaum, Jörg-Rüdiger Sack

Viagénie Inc. : André Cormier, Régis Desmeules, Guy Turcotte

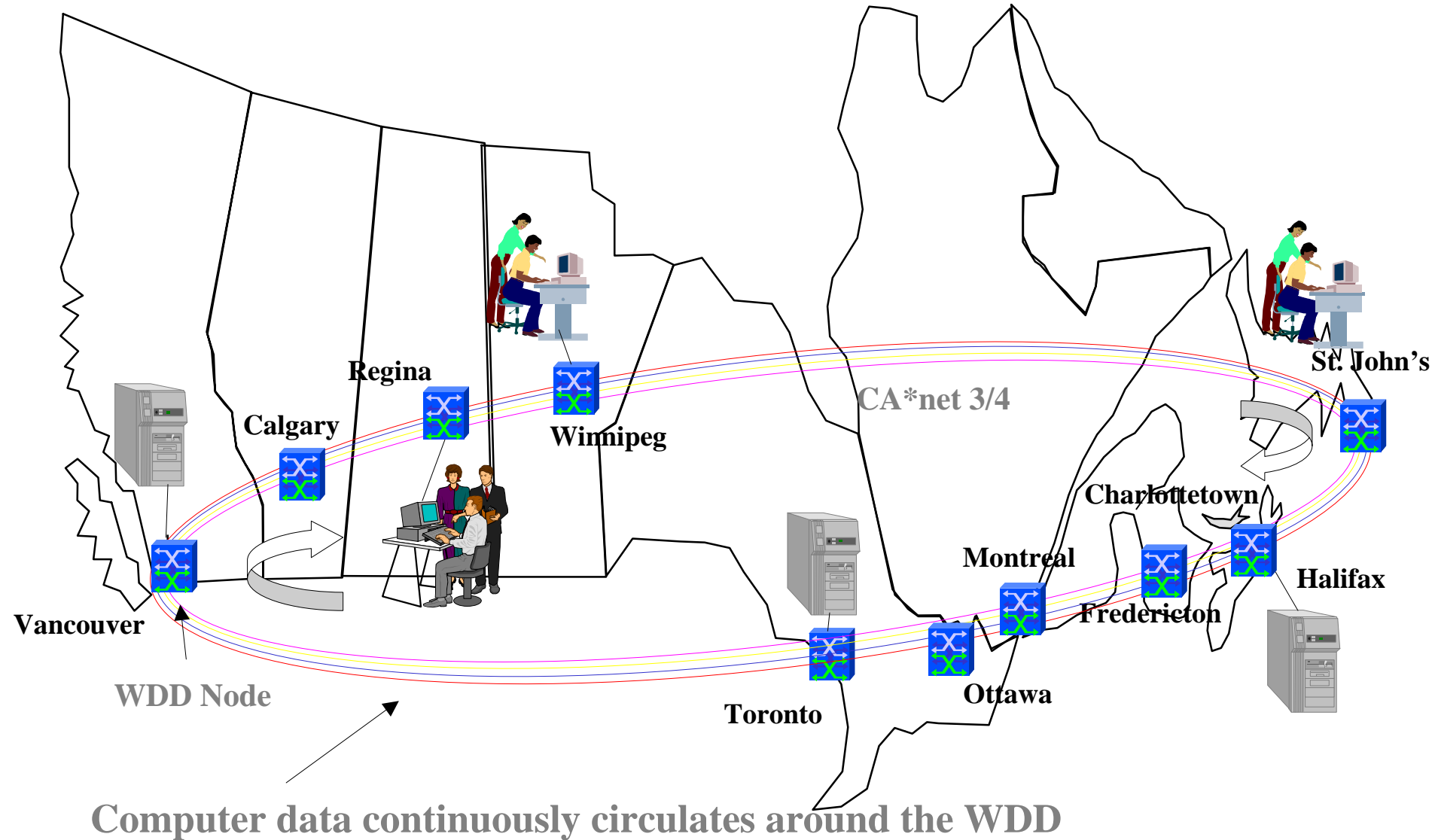
Why WDD ?

- making use of the expected proliferation of optical bandwidth
 - 8 OC-192 wavelengths across Canada have intrinsic storage capacity of approximately 1 GigaByte
- addressing issues with inter-processor communication in distributed computing environments and IP
 - TCP slow-start
 - big fat pipe problem
 - head-of-line blocking
 - n-squared connection problem

What is WDD ?

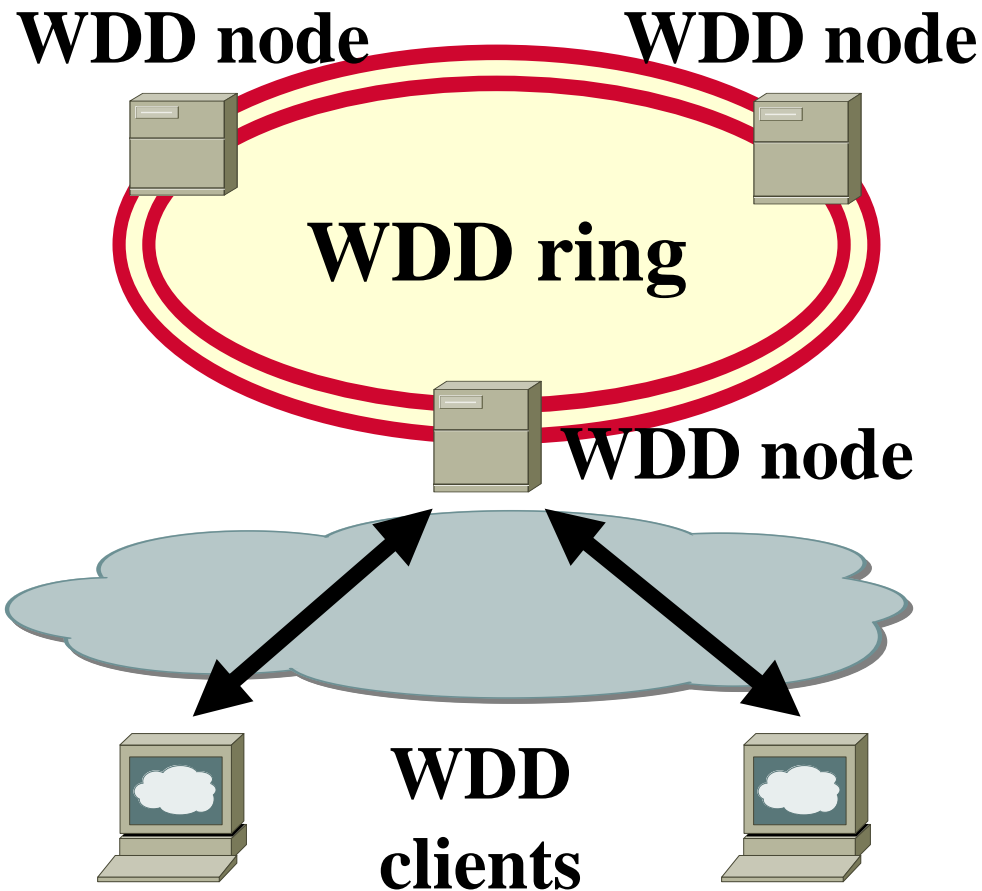
- not a hard disk neither a memory device on servers or computers
- the NETWORK is a storage device itself
- by configuring network in a special topology (ring), data can be stored and shared to a large number of users
- the storage potential in a Wavelength Disk Drive is determined by :
 - length of the network (circumference of the ring), and
 - bandwidth or number of wavelengths (diameter of the pipe)

Wavelength Disk Drives



WDD Main Concepts

Scalable Storage Ring



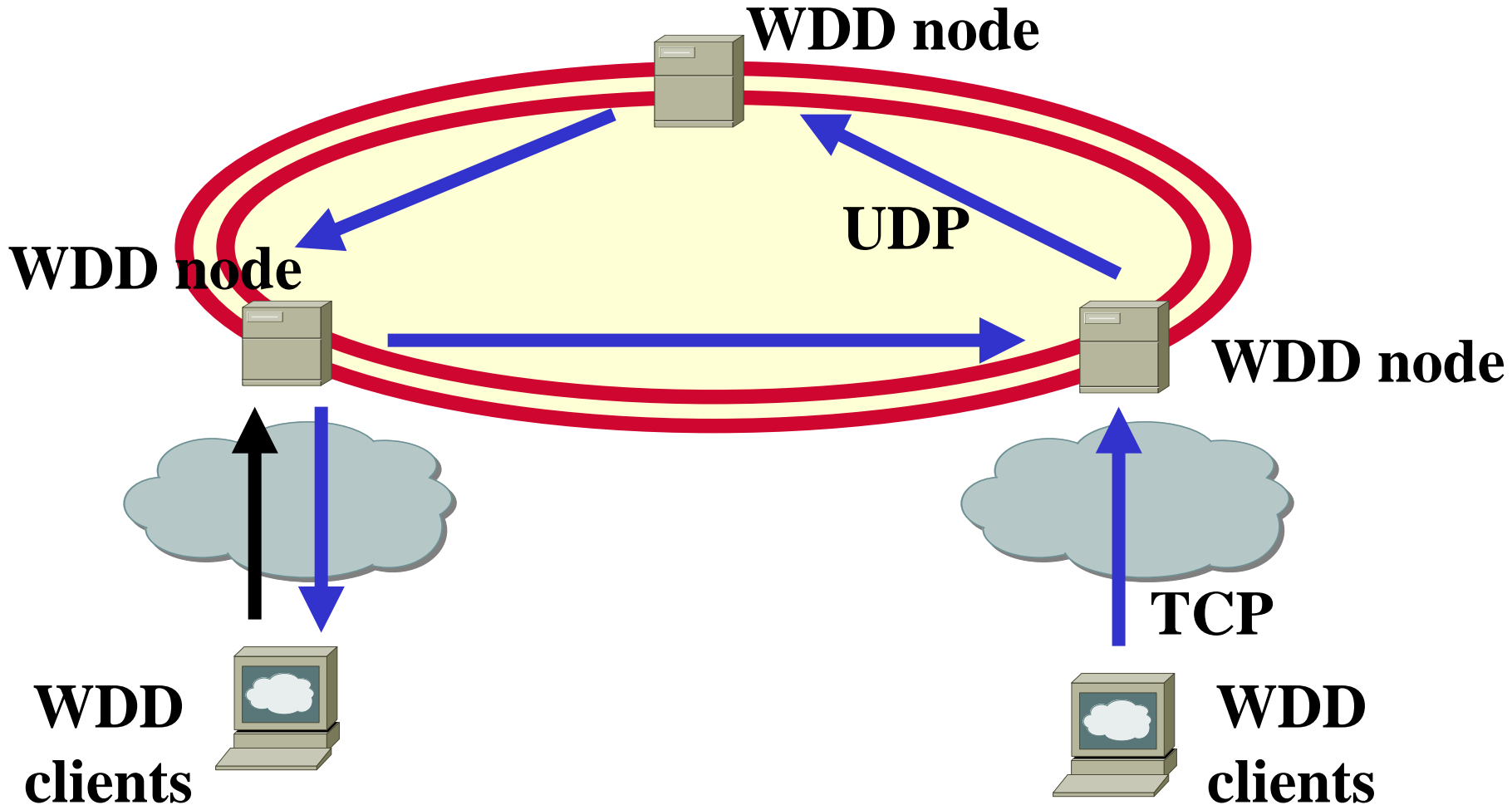
- a WDD system consists of :
 - a fibre optic ring
 - with two or more WDD Nodes
 - each WDD Node is attached to one or more application computers
- the system is scalable in :
 - the size of the ring
 - the number of WDD Nodes in the ring
 - the number of application computers supported by any one WDD Node

WDD Main Concepts

Content-Based Messaging

- a Wavelength Disk Drive system operates using content-based messaging, in which :
 - a message is created by a producer process, and
 - is placed in a pool of messages, until
 - it is removed from the message pool by a consumer process
- the producer need not know which consumer removes the message, and the consumer need not know which producer placed the message into the pool
- messages are managed in the pool by a description of their contents

WDD Operation





Phase I Project Goals

Working Prototype

- Nationwide Testbed Ring Deployment
- Client/Server API Definition via Application Requirements Analysis
- Evaluation and Validation via Application Usage
- Planning for Further Research

Phase I WDD Nodes

- WDD node is a computer with a GigE interface connected to an optical backbone (CA*net-3) through a core router (e.g. GSR)
- several nodes form a ring at the IP layer where data circulates (using UDP as transport)
- WDD daemon in each WDD node manages all data of the ring :
 - Entry/exit point of data submitted by WDD clients
 - Ensure the availability/removal of data circulating in the ring (checksum, data regeneration, deleting, expiration,..)

Phase I WDD Ring

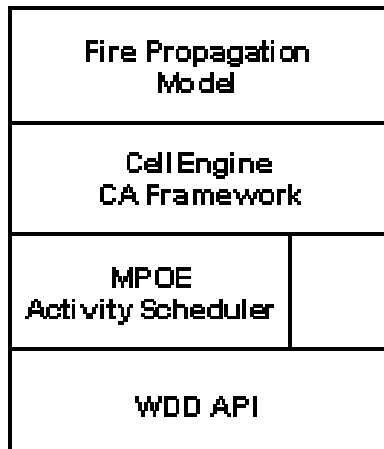
- WDD ring ensures data are circulated between all WDD nodes (UDP)
- messages are sent to the ring by WDD nodes
- upon requests from WDD clients, messages are retrieved by a WDD node and sent to the WDD client using TCP

Phase I WDD API

- a WDD API had been defined as control mechanism between WDD nodes and WDD clients for :
 - open/close WDD client/server session
 - create/delete instance of an application on ring
 - put/delete message to/from ring
 - request/cancel_request for message type from ring
 - processing of incoming WDD message in application
 - ASCII messages corresponding to error codes
- this API allows the development of WDD applications

Firesim

The Phase I Test Application



- the Firesim program uses a very simple fire propagation model
- Firesim inputs :
 - fuel map (land cover and vegetation description) :
 - water, rock, sand, swamp, wood[1,2,3,4]
 - wind intensity and direction
 - initial fires (intensity, location and time)
- Firesim calculations
 - burning state :
 - alive, warm[1,2], burning[1,2,3,4], burnt
 - output images; colours are dependent on burning state and fuel type



CellEngine - WDD Interaction

- input raster, initial block distribution
 - the Manager builds raster assignments (sub-raster blocks plus Firesim wind and initial fire controls) and places them onto the ring
 - the Manager is aware of the maximum amount of data it may place on the ring and only inserts the appropriate number of raster assignments
 - Workers place requests for raster assignments on the ring when idle
 - the Manager is informed by the ring whenever a Worker has consumed a raster assignment
 - this is done using the “Notify” capability of the API
 - this allows it to place additional assignments on the ring if any remain
- CA execution control
 - the Manager requests all status messages for the instance it initiated
 - Workers request all control messages for blocks they are processing

CellEngine - WDD Interaction (cont.)

- boundary data distribution
 - after each iteration of the CA, the Workers create messages containing the boundary information for each of the raster assignments they have taken on
 - correspondingly they also place requests for the boundary information for the neighbours of their raster assignments
- image data distribution
 - after every X iterations the Workers generate an image portion for each of their raster assignments which are placed on the ring as messages
 - the Manager places corresponding requests for the display messages

Firesim Application

Sample Output - No Wind



Firesim Application

Sample Output - No Wind (cont.)



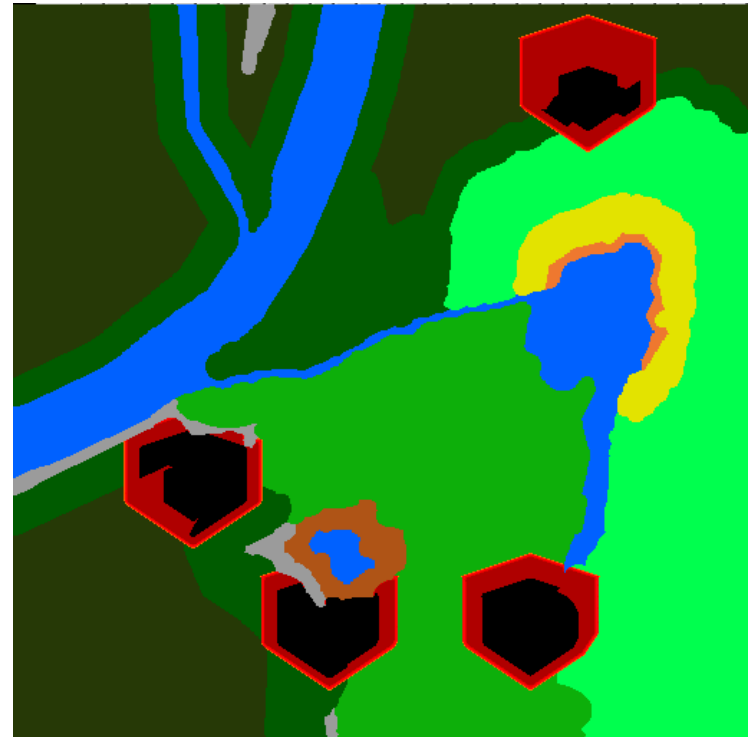
Firesim Application

Sample Output - Wind Value 25



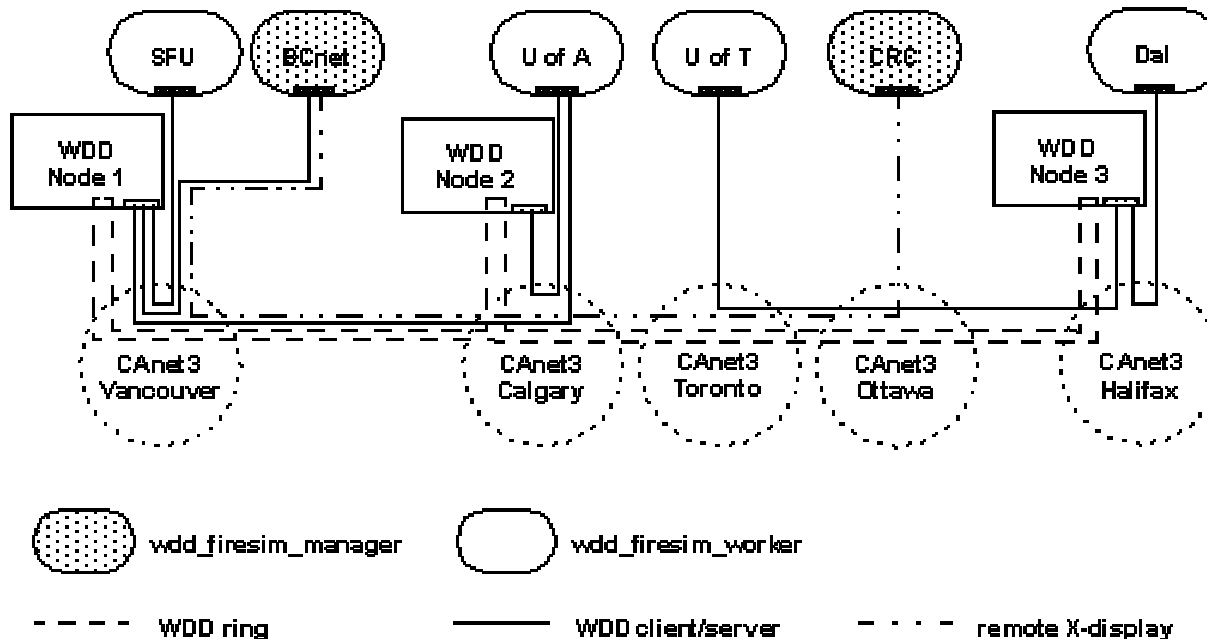
Firesim Application

Sample Output - Wind Value 25 (cont.)



WDD Phase I System Layout

- the Phase I WDD system consists of :
 - three WDD Nodes (Ring Servers), and
 - multiple application nodes (Ring Clients) defined by the two types of cellular automata engine components (manager and worker)



RESULTS : Phase I Goals Achieved

- Nationwide Testbed Ring Deployment
 - three node ring from Vancouver through Calgary to Halifax
- Client/Server API Definition via Application Requirements Analysis
 - the communication requirements of the parallel CA framework evolved a simple put/fetch protocol to capable, many-featured API
 - significant capabilities are still required, notably allocation and security
- Evaluation and Validation via Application Usage
 - test runs in various configurations of the Firesim application have validated all of the key WDD concepts
 - performance limitations of the Phase I system precluded performance evaluations; in effect, we know WDD “works” but not yet whether (or at least under which conditions) it “makes sense”
- Planning for Further Research
 - discussed later

RESULTS : Basic Concepts Proven

- Scalable Storage Ring
 - 2 and 3 node rings tested
 - variable number of clients connected to each ring server
 - ring testing applications as well as multiple instances of Firesim application concurrently executing
- Content Messaging Protocol
 - the defined WDD API allowed for the design and implementation of a fairly complex parallel application (CellEngine)
- Distributed Arbitration for Object Access/Consumption
- Efficient Long Haul Data Transfer in “Big Fat” pipes with UDP flows

RESULTS : Phase I WDD System Capability Analysis

- excessive processing demands on WDD nodes
 - only 1 processor in WDD nodes active (OS limitation)
 - WDD ring maintenance and client/server interaction (CMP processing) are all being done by the same CPU
 - GigE interface processing also being done by the same processor

possible solutions

- separate WDD ring processing and CMP processing functions to dedicated processors; possibly putting each function in different systems (this would add a new level in WDD client/server hierarchy)
- change to real-time, multi-processor OS in WDD nodes with one processor for GigE processing and one for WDD ring maintenance



RESULTS : Phase I WDD System Capability Analysis (cont.)

- excessive wire speed packet processing
 - all packets in ring pass through all WDD nodes
 - WDD ring diameter becomes restricted by capacity of link between router and WDD node

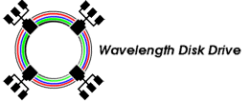
possible solutions

- identify messages in WDD ring with unique IP addresses (10.x.x.x/y) and use routers to classify and forward packets at wire speed
- WDD nodes express interest in IP addresses corresponding to only the messages they “need”
- this will allow a higher capacity ring (i.e. OC-192) with only GigE connections between routers and WDD nodes



RESULTS : CMP and Distributed CMP Paradigms

- the WDD project started with a concept for a novel use of optical networks
- a basic put/fetch protocol to use the lab prototype WDD ring was devised
- in considering the CellEngine and Firesim application, a more robust and complete WDD API was developed; this API has also been dubbed CMP or Content Messaging Protocol
- although our Phase I experiment is of a distributed nature, CMP does not require this; as such this project has spawned research into local CMP (within a single computing system or LAN) and distributed CMP (with multiple collaborating CMP servers)
- WDD is in effect the first technology being explored for implementing a distributed CMP system; other options for implementing distributed CMP systems will be explored including WDD in conjunction with other technologies



NEXT : Phase II Project Goals

- Migrate to CA*net 4; Full CA*net4 coverage
- Improved Performance and Capacity
- Enhanced Robustness and Reliability
- Multiple, Real-world Application Evaluation

NEXT : WDD Research

- WDD Node Update
 - separate WDD Ring functions from CMP functions
 - separate packet classification and forwarding from WDD Ring maintenance and use routers for wire-speed functions
- WDD Ring Update, Testbed Enlargement
 - deploy WDD Nodes in all provinces
 - deploy WDD Node at StarTap in Chicago
 - deploy WDD Nodes at other StarLight partners creating a truly global WDD Ring

NEXT : WDD Research (cont.)

- Enhance CMP (addressing security and allocation issues as well as adding capabilities to support other application requirements) and explore submitting to IETF or GGF
- Extend CMP support to other platforms and architectures (Linux, clusters, SMP, etc.)
- Integrate with existing Grid toolkits (e.g. Globus and/or Legion)



NEXT : Possible Phase II Application Research Areas

- Sudbury Neutrino Observatory, SNOMAN
- ALTA Project, cosmic ray detection and analysis
- TeraScale Computing/Data Mining/Visualization on StarLight
- Distributed storage with YottaYotta
- Mesoscale weather modeling, MC2 from Environment Canada
- Ice Tracking at Carleton University's PARADIGM Research Group



WHO : Computing Resource and Site Providers (many thanks!)

- BC*net
- Communications Research Centre, BADLAB
- Dalhousie University, Oceanography Department
- Silicon Graphics Inc.
- Simon Fraser University, Centre for Experimental & Constructive Mathematics
- University of Alberta, MACI Facility
- University of Toronto, Chemistry Department



Questions.....

<http://www.canet3.net/wdd>