CRAY
THE SUPERCOMPUTER COMPANY

Cray Operating System Plans and Status

May 2010

Charlie Carroll

Notice of Funding

This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

Cray Operating Systems and I/O

- Compute node OS
  - CNL
- Service node OS
  - Supports all compute nodes
- File systems
  - Lustre
  - DVS (Data Virtualization Service)
- Networking
  - HSN: Gemini drivers
  - TCP/IP
  - HSN: Portals
- Operating system services
  - Node Health Checker
  - Core specialization
  - DSL support
  - Cluster Compatibility Mode
- System management
  - CMS (Cray Management Services)
  - ALPS (Application-Level Placement Scheduler)
    - Interfaces to batch schedulers
  - Command interface

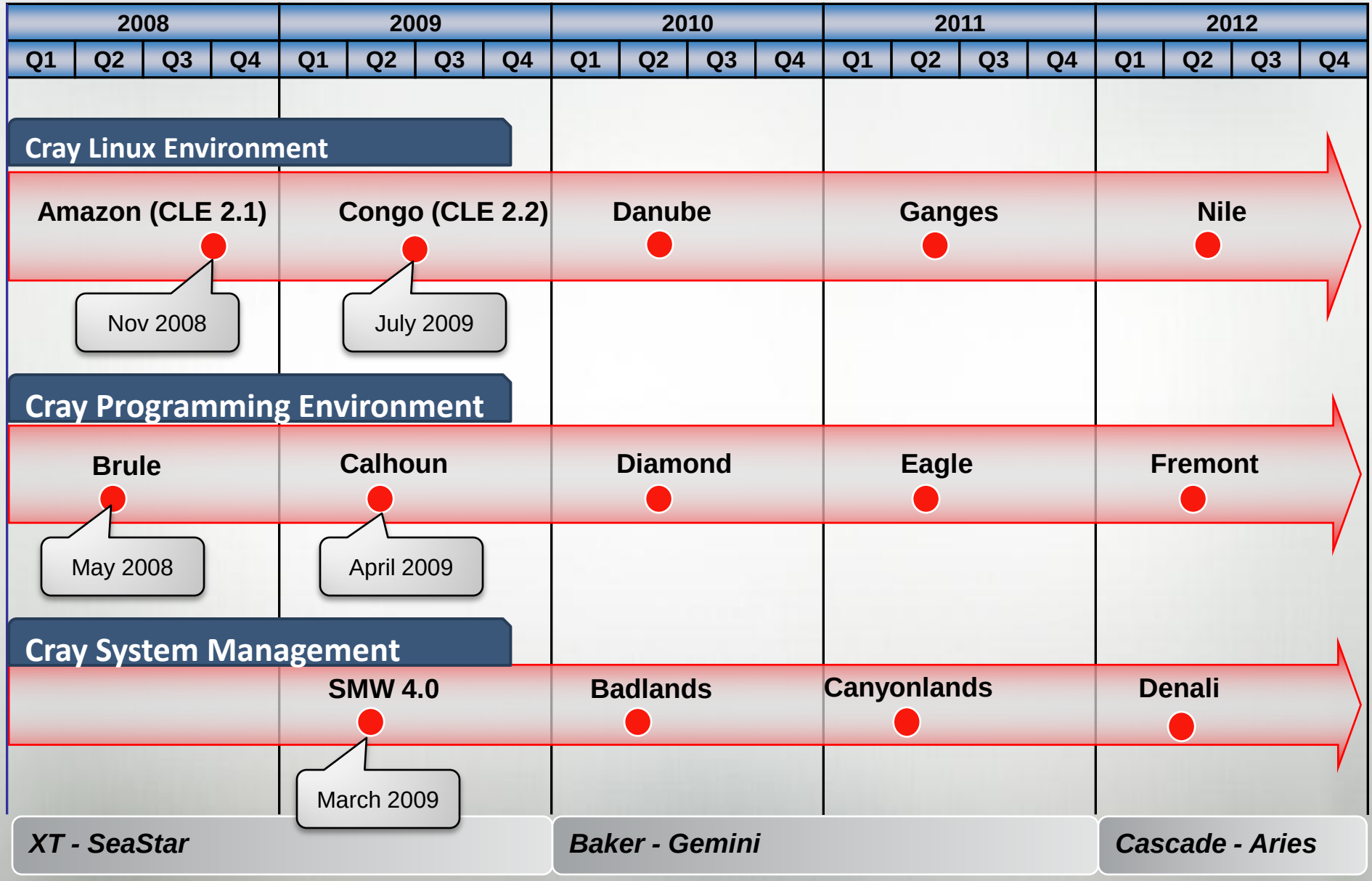
Cray Operating Systems Focus

- Performance
  - Maximize compute cycles delivered to applications while also providing necessary services
    - Lightweight operating system on compute node
    - Standard Linux environment on service nodes
  - Optimize network performance through close interaction with hardware
- Stability and Resiliency
  - Correct defects which impact stability
  - Implement features to increase system and application robustness
- Scalability
  - Scale to large system sizes without sacrificing stability
  - Provide better system management tools to manage more complicated systems

OSIO Accomplishments of the Past Year

- CLE 2.2
  - DVS: load balancing and cluster parallel mode
  - Dynamic Shared Library (DSL) support
- CLE 3.0 and SMW 5.0
  - XT6 (Magny-Cours + SeaStar) support
  - SLES11 and Lustre 1.8.1
  - DVS stripe parallel mode
- CLE 3.1 and SMW 5.1
  - Gemini support
  - Core specialization
  - Cluster Compatibility Mode (CCM)
  - DVS failover
- Software Mean Time to Interrupt (SMTTI) up to ~2500 hours

Cray Software - At a glance



Danube: Gemini—A New High-Speed Network

- Replaces SeaStar and Portals
- First shipments in 2H10
- New high-speed network software stack with far-reaching implications
  - Portals replaced with two new APIs
    - User-level Gemini Network Interface (uGNI)
    - Distributed memory application interface (DMAPP)
  - Better error handling
  - Less done in software
- Better performance: ~1.7us ping-pong latency
- Link resiliency
  - Adaptive routing: multiple paths to the same destination
  - System able to survive link outages
  - Warm swap: reroute; quiesce; swap; activate

Danube: Core Specialization

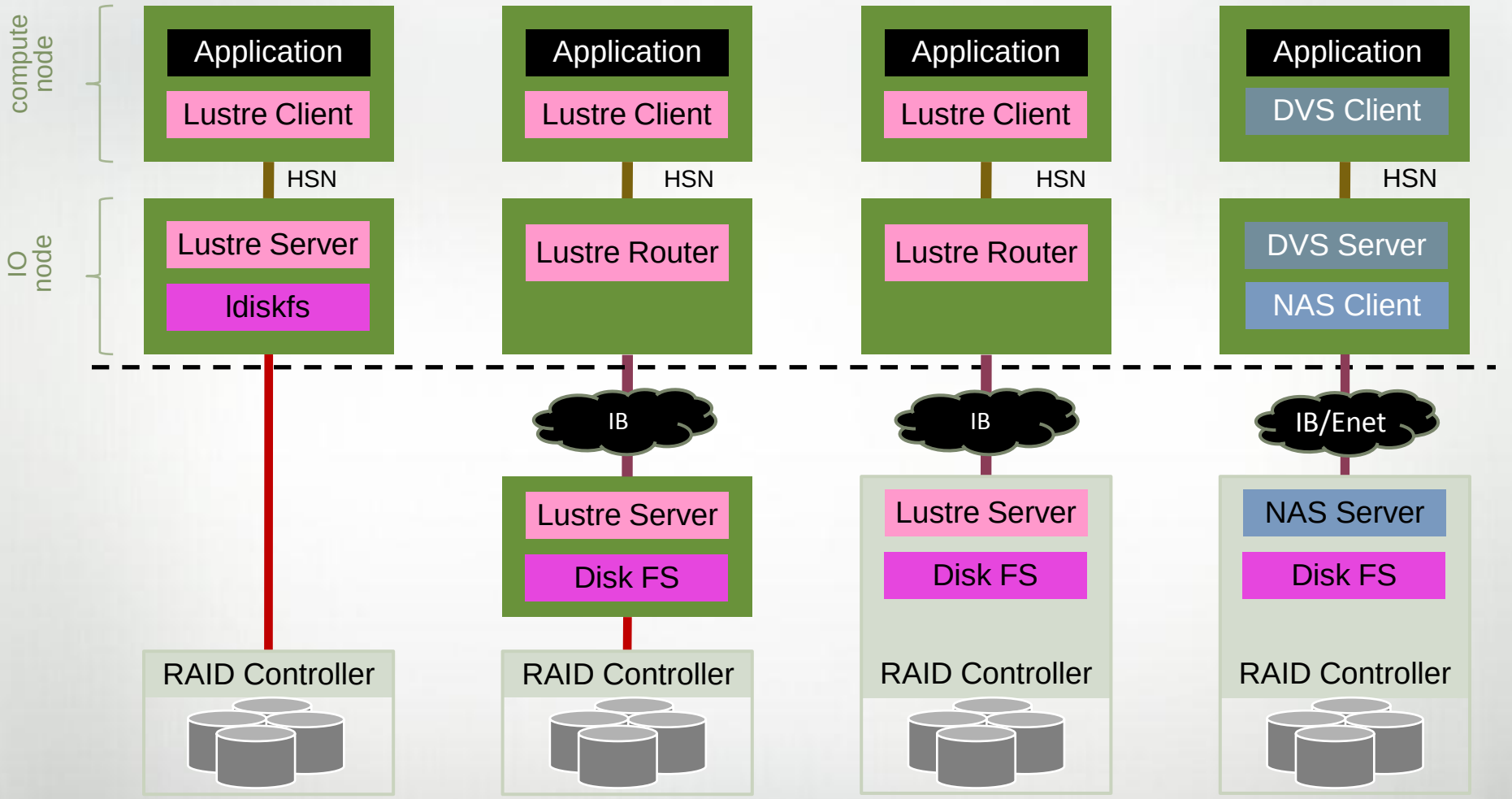
- Benefit
  - Can improve performance by reducing noise on compute cores
  - Moves overhead (interrupts, daemon execution) to a single core
- Rearranges existing work
  - Without core specialization: overhead affects every core
  - With core specialization: overhead is confined, giving application exclusive access to remaining cores
- Helps some applications, hurts others
  - POP 2.0.1 on 8K cores on XT5: 23% improvement
  - Larger jobs see larger benefit
- Optional on a job-by-job basis
  - By default core specialization is "off"
  - Launch switch enables this feature

Danube: Cluster Compatibility Mode (CCM)

- Provides the runtime environment on compute nodes expected by ISV applications
- Dynamically allocates and configures compute nodes at job start
  - Nodes are not permanently dedicated to CCM
  - Any compute node can be used
  - Allocated like any other batch job (on demand)
- MPI and third-party MPI runs over TCP/IP over high-speed network
- Supports standard services: ssh, rsh, nscd, ldap
- Complete root file system on the compute nodes
  - Built on top of the Dynamic Shared Libraries (DSL) environment
- Apps run under CCM: Abaqus, Matlab, Castep, Discoverer, Dmo13, Mesodyn, Ensight and more

Under CCM, everything the application can “see” is like a standard Linux cluster: Linux OS, x86 processor, and MPI

Cray I/O Models



Direct-Attach Lustre

External Lustre

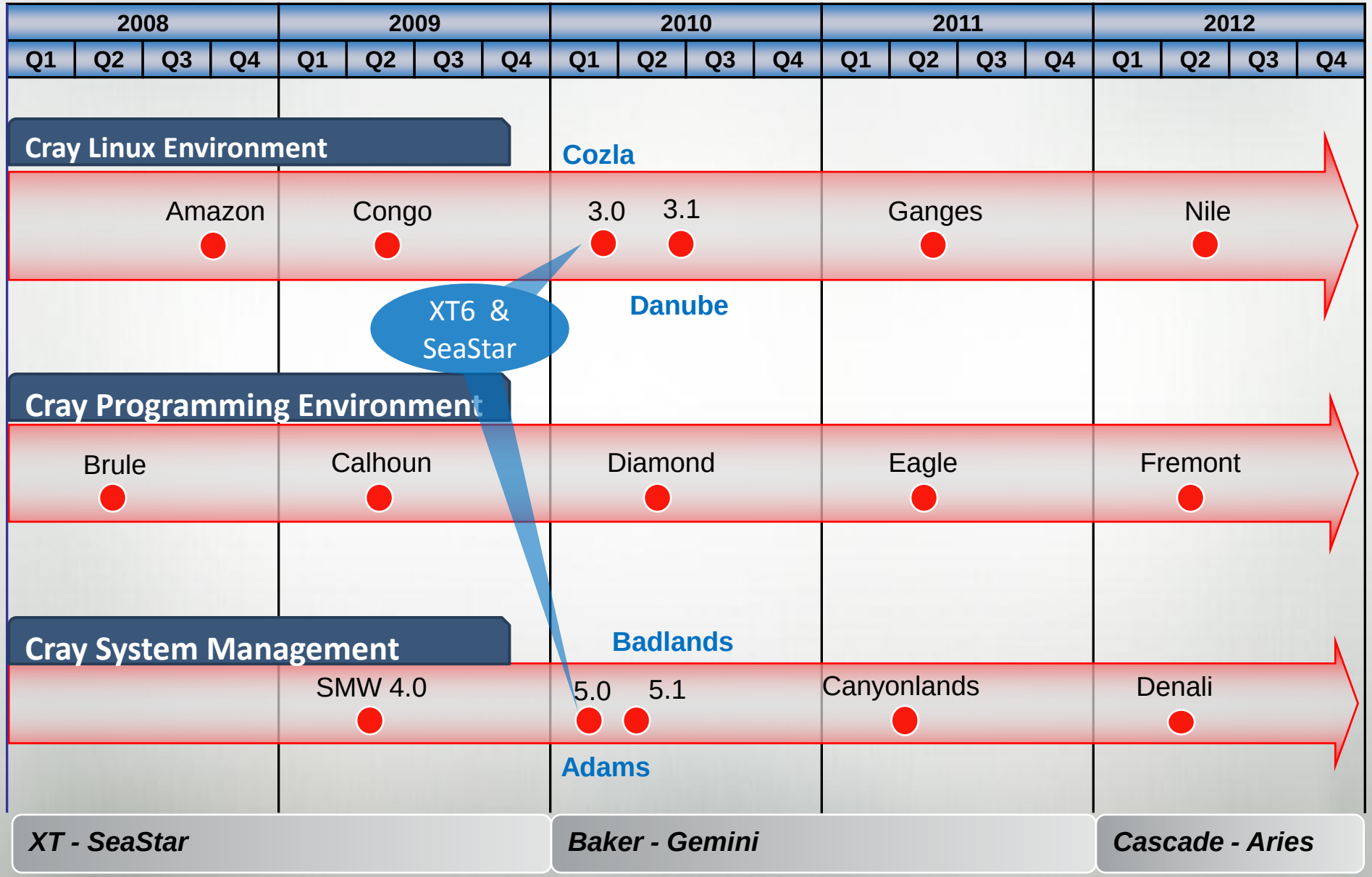
Lustre Appliance

Alternate External File Systems
(GPFS, Panasas, NFS)

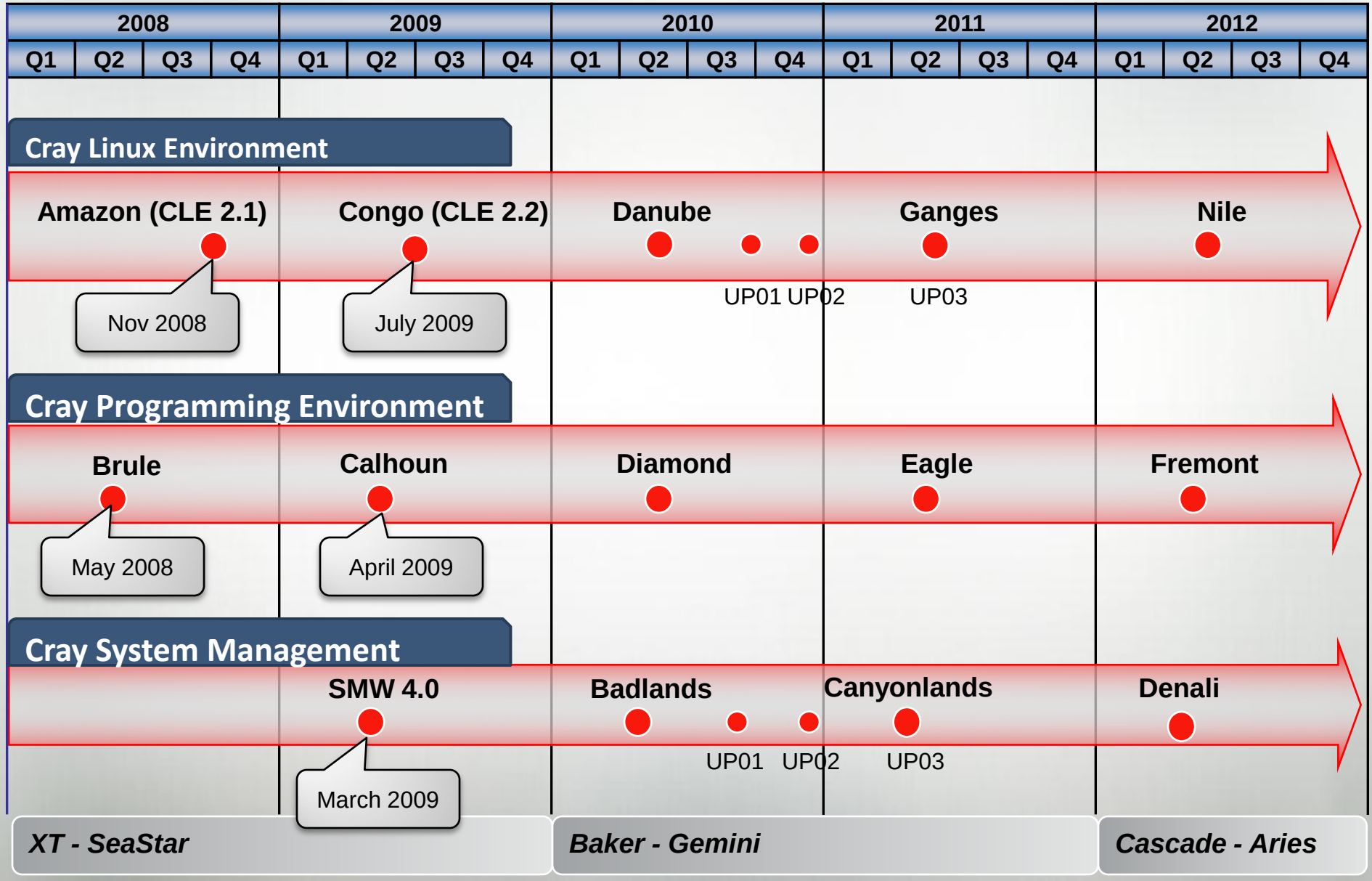
Danube: File System Technologies

- Lustre 1.8
  - Failover improvements
    - Version Based Recovery
    - Imperative recovery
  - OSS cache
  - Adaptive timeouts
  - OST pools
- DVS (Data Virtualization Service)
  - Stripe parallel mode
  - Failover and failback

Cray Software – XT6 Releases



Cray Software – Danube Updates



CLE 3.1 UP01

- RSIP scaling
- Repurposed Compute Nodes (Moab/Torque only)
  - Configure compute node hardware with service node software
  - Login nodes, MOM nodes, DSL servers
- Lustre 1.8.2
- Performance improvements to Gemini stack
  - Shared small message buffers

- Blue = Defining feature
- Black = Target feature

CLE 3.1 UP02

- XT4 and XT5 support
- CCM: ISV application acceleration
  - Leverages part of the OFED stack to support multiple third-party MPIs directly over the Gemini-based high-speed network
- DVS-Panasas support
- Checkpoint / restart
- Lustre 1.8.3

Release Support

	XT3	XT4	XT5	XT6	Baker	Gemini Upgrade
CLE 2.2	Yes	Yes	Yes			
CLE 3.0				Yes		
CLE 3.1				Yes	Yes	
CLE 3.1 UP01				Yes	Yes	Yes
CLE 3.1 UP02		Yes	Yes	Yes	Yes	Yes
CLE 3.1 UP03		Yes	Yes	Yes	Yes	Yes
Ganges					Yes	Yes

Summary

- Cray is about to release the software stack to support our new interconnect, new SIO blade and new processor
  - CLE 3.1 (aka Danube), SMW 5.1 in June 2010
- Updates to CLE 3.1 and SMW 5.1 will include features
  - CLE 3.1 UP02 will bring Danube support to XT5s and XT4s
- Ganges (Jun 2010) will support Interlagos
- Software quality continues to improve


CRAY
THE SUPERCOMPUTER COMPANY