PDL Packet Spring Update

NEWSLETTER ON THE PARALLEL DATA LABORATORY • SPRING 2005

http://www.pdl.cmu.edu/

CONSORTIUM MEMBERS

EMC
EqualLogic
Hewlett-Packard
Hitachi
Hitachi Global Storage
IBM
Intel
Microsoft Research
Network Appliance
Oracle
Panasas
Seagate
Sun Microsystems

CONTENTS

THE PDL PACKET

EDITOR
Joan Digney

CONTACT
Greg Ganger
PDL Director

Karen Lindenfelser
PDL Business Administrator

The Parallel Data Laboratory
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
TEL 412-268-6716
FAX 412-268-3010

http://www.pdl.cmu.edu/Publications/

SELECTED RECENT PUBLICATIONS

Comparison-Based File Server Verification

Tan, Wong, Strunk & Ganger

USENIX '05 Annual Technical Conference, April 10-15, 2005. Anaheim, CA.

Comparison-based server verification involves testing a server by comparing its responses to those of a reference server. An intermediary, called a "server Tee," interposes between clients and the reference server, synchronizes the system-under-test (SUT) to match the reference server's state, duplicates each request for the SUT, and compares each pair of responses to identify any discrepancies. The result is a detailed view into any differences in how the SUT satisfies the client-server protocol specification, which can be invaluable in debugging servers, achieving bug compatibility, and isolating performance differences. This paper introduces, develops, and illustrates the use of comparison-based server verification. As a concrete example, it describes a NFSv3 Tee and reports on its use in identifying interesting differences in several production NFS servers and in debugging a prototype NFS server. These experiences confirm that comparison-based server verification can be a useful tool for server implementors.

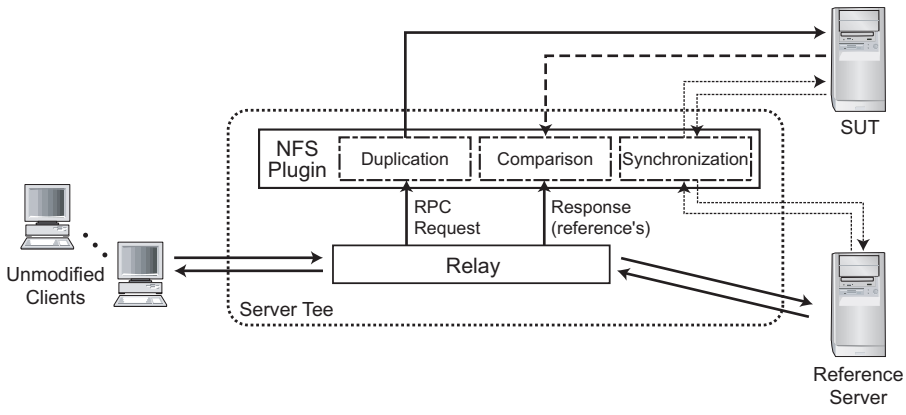
Challenges in Building a Two-Tiered Learning Architecture for Disk Layout

Salmon, Thereska, Soules, Strunk & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-04-109, August, 2004.

Choosing the correct settings for large systems can be a daunting task. The performance of the system is often heavily dependent upon these settings, and the "correct" settings are often closely coupled with the workload. System designers usually resort to using a set of heuristic approaches that are known to work well in some cases. However, hand-combining these heuristics is painstaking and fragile. We propose a two-tiered architecture that makes this combination transparent and robust, and describe an application of the architecture to the problem of disk layout optimization. This two-tiered architecture

… *continued on pg. 2*



Software architecture of an NFS Tee. To minimize potential impact on clients, we separate the relaying functionality from the other three primary Tee functions (which contain the vast majority of the code). One or more NFS plug-ins can be dynamically initiated to compare a SUT to the reference server with which clients are interacting.

PDL NEWS

http://www.pdl.cmu.edu/News/

March 2005
Matthew Wachs awarded NDSEG Fellowship

Congratulations to Matthew Wachs, who has been selected to receive a 2005-2006 National Defense Science and Engineering Graduate (NDSEG) Fellowship. This fellowship is sponsored by the Department of Defense through the Air Force Office of Scientific Research (AFOSR), the Office of Naval Research (ONR), the Army Research Office (ARO), and the High Performance Computing Modernization Program, and is administered by the American Society for Engineering Education (ASEE).

March 2005
Dawn Song Receives NSF CAREER AWARD

Dawn Song, assistant professor of Electrical and Computer Engineering and Computer Science, has received a National Science Foundation CAREER Award for her research proposal, "Toward Exterminating Large Scale Internet Attacks." The award "recognizes and supports the early career-development activities of those teacher-scholars who are most likely to become the academic leaders of the 21st century."

–with info from CMU's 8 1/2 x 11 News

January 2005
Anastassia Ailamaki selected as a Sloan Research Fellow

We are extremely happy to announce the CMU 2005 winners of a Sloan Research Fellowship: Natassa Ailamaki, Karl Crary and Anupam Gupta. A Sloan Fellowship is a prestigious award intended to enhance the careers of the very best young faculty members in specified fields of science. Currently a total of 116 fellowships are awarded annually in seven fields: chemistry, computational and evolutionary molecular biology, computer science, economics, mathematics, neuroscience, and physics.

See more about the Sloan fellowships at http://www.sloan.org/programs/scitech_fellowships.shtml

December 2004
Outstanding Researchers

Professor David O'Hallaron and Research Engineer Volkan Akcelik have been awarded the Outstanding Research Award for their work on the Quake Project by the College of Engineering in its CIT Faculty Awards for

… continued on pg. 4

RECENT PUBLICATIONS

… continued from pg. 1

consists of a set of independent heuristics, and an adaptive method of combining them. However, building such a system has proved to be more difficult than expected. Each heuristic depends heavily on decisions from other heuristics, making it difficult to break the problem into smaller pieces. This paper outlines our approaches and how they have worked, discusses the biggest challenges in building the system, and mentions additional possible solutions. Whether this problem is solvable is still open to debate, but the experiences reported provide a cautionary tale; system policy automation is complex and difficult.

What If You Could Ask "What-if"? (Simplifying Large System Administration)

Thereska, Narayanan & Ganger

Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-05-101, February 2005.

Today, management and tuning questions are approached using if/then rules of thumb. This reactive approach requires expertise regarding of system behavior, making it difficult to deal with unforeseen uses of a system's resources and leading to system unpredictability and large system management overheads. We propose a "what if" approach that allows interactive exploration of the effects of system changes, thus converting complex tuning problem into simpler search problems. Through two concrete management problems, automating system upgrades and deciding on service migrations, we identify system design changes that enable a system to answer "what if" questions about itself.

On Hierarchical Routing in Doubling Metrics

Gupta, Maggs & Zhou

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-04-106, December 2004.

We study the problem of routing in doubling metrics, and show how to perform hierarchical routing in such metrics with small stretch and compact routing tables (i.e., with a small amount of routing information stored at each vertex). We say that a metric *(X,d)* has *doubling dimension dim(X)* at most a if every set of diameter D can be covered by 2^{α} sets of diameter *D/2*. (A *doubling metric* is one whose doubling dimension *dim(X)* is a constant.) For a connected graph G, whose shortest path distances d_G induce the doubling metric (X, d_G), we show how to perform $(1 + \tau)$-stretch routing on G for any $0 < \tau \leq 1$ with routing tables of size at most $(\alpha/\tau)^{O(\alpha)} \log\Delta \log\delta$ bits with only $(\alpha/\tau)^{O(\alpha)} \log\Delta$ entries, where Δ is the diameter of G and δ is the maximum degree of G. Hence the number of routing table entries is just $\tau^{-O(1)} \log\Delta$ for doubling metrics. These results extend and improve on those of Talwar (2004).

… continued on pg. 3

… continued from pg. 2

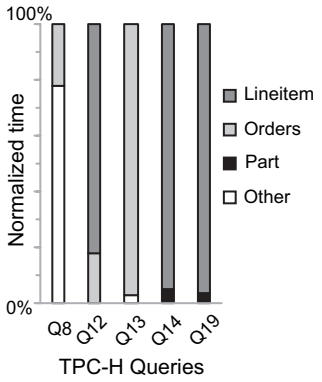
QPipe: A Simultaneously Pipelined Relational Query Engine

Harizopoulos, Shkapenyuk & Ailamaki

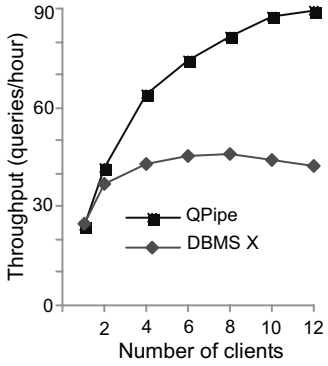
SIGMOD 2005, June 14-16, 2005, Baltimore, Maryland, USA.

Relational DBMS typically execute concurrent queries independently by invoking a set of operator instances for each query. To exploit common data retrievals and computation in concurrent queries, researchers have proposed a wealth of techniques, ranging from buffering disk pages to constructing materialized views and optimizing multiple queries. The ideas proposed, however, are inherently limited by the query-centric philosophy of modern engine designs. Ideally, the query engine should proactively coordinate same-operator execution among concurrent queries, thereby exploiting common accesses to memory and disks as well as common intermediate result computation.

This paper introduces on-demand simultaneous pipelining (OSP), a novel query evaluation paradigm for maximizing data and work sharing across concurrent queries at execution time. OSP enables proactive, dynamic operator sharing by pipelining the operator's output simultaneously to multiple parent nodes. This paper also introduces QPipe, a new operator-centric relational engine that effortlessly supports OSP. Each relational operator is encapsulated in a micro-engine serving query tasks from a queue, naturally exploiting all data and work sharing opportunities. Evaluation of QPipe built on top of BerkeleyDB shows that QPipe achieves a 2x speedup over a commercial DBMS when running a workload consisting of TPC-H queries.



Time breakdown for five TPC-H queries. Each component shows time spent reading a TPC-H table.



Throughput for one to twelve concurrent clients running TPC-H queries on DBMS X and QPipe.

… continued on pg. 5

DISSERTATION ABSTRACT: Cluster Scheduling for Explicitly-Speculative Tasks

David Petrou

Carnegie Mellon University, Dept. ECE Ph.D. Dissertation CMU-PDL-04-112. December 2004.

A process scheduler on a shared cluster, grid, or supercomputer that is informed which submitted tasks are possibly unneeded speculative tasks can use this knowledge to better support increasingly prevalent user work habits, lowering user-visible response time, lowering user costs, and increasing resource provider revenue.

Large-scale computing often consists of many speculative tasks (tasks that may be canceled) to test hypotheses, search for insights, and review potentially finished products. For example, speculative tasks are issued by bioinformaticists comparing dna sequences, computer graphics artists rendering scenes, and computer researchers studying caching. This behavior—exploratory searches and parameter studies, made more common by the cost effectiveness of cluster computing—on existing schedulers without speculative task support results in a mismatch of goals and suboptimal scheduling. Users wish to reduce their time waiting for needed task output and the amount they will be charged for unneeded speculation, making it unclear to the user how many speculative tasks they should submit.

This thesis introduces 'batchactive' scheduling (combining batch and interactive characteristics) to exploit the inherent speculation in common application scenarios. With a batchactive scheduler, users submit explicitly labeled batches of speculative tasks exploring ambitious lines of inquiry, and users interactively request task outputs when these outputs are found to be needed. After receiving and considering an output for some time, a user decides whether to request more outputs, cancel tasks, or disclose new speculative tasks. Users are encouraged to disclose more computation because batchactive scheduling intelligently prioritizes among speculative and non-speculative tasks, providing good wait-time-based metrics, and because batchactive scheduling employs an incentive pricing mechanism which charges for only requested task outputs (i.e., unneeded speculative tasks are not charged), providing better cost-based metrics for users. These aspects can lead to higher billed serv-

… continued on pg. 4

… continued from pg. 3

er utilization, encouraging batchactive adoption by resource providers organized as either cost- or profit-centers.

Not all tasks are equal—only tasks whose outputs users eventually desire matter—leading me to introduce the 'visible response time' metric which accrues for each task in a batch of potentially speculative tasks when the user needs its output, not when the entire batch was submitted, and the batchactive pricing mechanism of charging for only needed tasks, which encourages users to disclosure future work while remaining resilient to abuse. I argue that the existence of user think times, user away periods, and server idle time makes batchactive scheduling applicable to today's systems.

I study the behavior of speculative and non-speculative scheduling using a highly-parameterizable discrete-event simulator of user and task behavior based on important application scenarios. I contribute this simulator to the community for further scheduling research.

For example, over a broad range of realistic simulated user behavior and task characteristics, I show that under a batchactive scheduler visible response time is improved by at least a factor of two for 20% of the Simulations. A batchactive scheduler which favors users who historically have desired a greater fraction of tasks that they speculatively disclosed provides additional performance and is resilient to a denial-of-service. Another result is that visible response time can be improved while increasing the throughput of tasks whose outputs were Desired. Under some situations, user costs decrease while server revenue increases. A related result is that more users can be supported and greater server revenue generated while achieving the same mean visible response time. A comparison against an impractical batchactive scheduler shows that the easily implementable two tiered batchactive schedulers, out of all batchactive schedulers, provide most of the potential performance gains. Finally, I demonstrate deployment feasibility by describing how to integrate a batchactive scheduler with a popular clustering system.

Raja takes a break from Retreat '04 proceedings.

DISSERTATION ABSTRACT:
Efficient, Scalable Consistency for Highly Fault-tolerant Storage

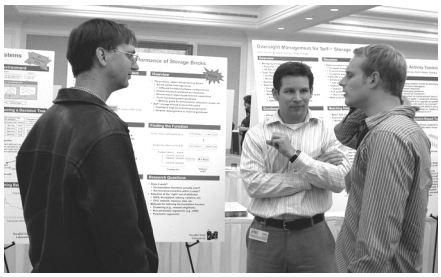
Garth Goodson

Carnegie Mellon University Ph.D Dissertation. CMU-PDL-04-111, August 2004.

Fault-tolerant storage systems spread data redundantly across a set of storage-nodes in an effort to preserve and provide access to data despite failures. One difficulty created by this architecture is the need for a consistent view, across storage-nodes, of the most recent update. Such consistency is made difficult by concurrent updates, partial updates made by clients that fail, and failures of storage-nodes.

This thesis demonstrates a novel approach to achieving scalable, highly fault-tolerant storage systems by leveraging a set of efficient and scalable, strong consistency protocols enabled by storage-node versioning. Versions maintained by storage-nodes can be used to provide consistency, without the need for central serialization, and despite concurrency. Since versions are maintained for every update, even if a client fails part way through an update, concurrency exists during an update, the latest complete version of the data-item being accessed still exists in the system—it does not get destroyed by subsequent updates. Additionally, versioning enables the use of optimistic protocols.

This thesis develops a set of consistency protocols appropriate for constructing block-based storage and metadata services. The block-based storage protocol is made space efficient through the use of erasure codes and made scalable by offloading work from the storage-nodes to the clients. The metadata service is made scalable by avoiding the high costs associated with agreement algorithms and by utilizing threshold voting quorums. Fault-tolerance is achieved by developing each protocol in a hybrid storage-node faultmodel (a mix of Byzantine and crash storage-nodes can be tolerated), capable of tolerating crash or Byzantine clients, and utilizing asynchronous communication.



Brandon Salmon and Mike Mesnier discuss their research with Richard New of HGST.

THESIS PROPOSAL:
Automating Attribute Assignment Using Context to Assist in Personal File Retrieval

Craig Soules

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Feb. 25, 2005.

Context information consists of both how a user perceives a piece of data, as well as how the user perceives the connection between data. A recent user study showed that most users use context information to locate their data during a search, however most organizational systems and search tools do not take this into account. In my work, I propose schemes to automatically gather context information from user activity and use that to generate file attributes for use in organization and search tools. By increasing the number of available attributes, I believe that such tools can be made more effective.

PDL NEWS

… continued from pg. 2

2004–2005. The Quake project is joint effort by the Dept of Civil and Environmental Engineering and the School of Computer Science at Carnegie Mellon University. Its goal is to develop the capability for predicting, by computer simulation, the ground motion of large basins during strong earthquakes, and to use this capability to study the seismic response of the Greater Los Angeles Basin. More information on the Quake project may be found at http://www-2.cs.cmu.edu/~quake/

–with info from CMU's 8 1/2 x 11 News

November 2004
Pittsburgh – The Epicenter of Storage Innovation

A reception highlighting Pittsburgh as the Epicenter of Storage Innovation is being held on Thursday, November 11, as a part of Supercomputing 2004, hosted by Pittsburgh's Storage Innovators. These innovators include the Data Storage Systems Center at Carnegie Mellon University, Intel Research Pittsburgh, Network Appliance, Panasas, the Parallel Data Laboratory at Carnegie Mellon University, the Pittsburgh Digital Greenhouse, the Pittsburgh Supercomputing Center and Seagate Research.

SC04 in Pittsburgh marks the first year that storage will get explicit recognition in high-performance computing. Pittsburgh has long been a leader in storage innovations that have spread out from here and left their mark on the entire industry. Dean Jim Morris of Carnegie Mellon University, former Dean of the School of Computer Science and current Dean of Carnegie Mellon University West Coast Campus will provide a short history of storage innovation in Pittsburgh. This history stretches from the days of the Information Technology Center (ITC) in the early 80s, and the creation of the Andrew File System (AFS), which later became a product from Transarc Corporation and then from IBM. AFS in turn inspired Coda, one basis for the distributed storage work now continuing at Intel Labs–Pittsburgh; led to Multi-Resident AFS (MR-AFS), still in production use at the Pittsburgh Supercomputing Center (PSC); and inspired the global file system developed by Spinnaker Networks (now Network Appliance, the market leader in networked storage). The Data Storage Systems Center (DSSC) was founded in 1990 to study advanced magnetic recording, inventing some of the basic technology and training many of the technical innovators who are still increasing the capacity of disk drive storage at the amazing rate of 50% per year. The DSSC provided early support for the Parallel Data Laboratory (PDL), which today leads the academic community in research into storage systems and originated the technology for Network Attached Secure Disks (NASD), a core component of the products currently being developed by Panasas. The DSSC also led directly to the founding of Seagate Research, which provides central R&D for the world's largest disk drive maker right here in Pittsburgh. The Pittsburgh Digital Greenhouse (PDG) supports commercialization and technology transfer, including storage technology, in the greater Pittsburgh region.

–with info from SC04 technical program notes

October 2004
CMU & PDL Host Posix Extensions Workshop

Carnegie Mellon University and the Parallel Data Lab hosted a Posix Extensions Workshop on November 8. The goal for the workshop was to achieve a well accepted by industry POSIX I/O API extension, or set of extensions, to make the POSIX I/O API more friendly to HPC, clustering, parallelism, and high concurrency applications. The meeting was held in conjunction with SuperComputing 2004 and covered the initial mechanics for how the POSIX API is to be extended, ideas for extensions, and the formation of a plan of attack, organization, and mapping of next steps.

September 2004
Spiros Papadimitriou Named a Siebel Scholar

Congratulations to Spiros Papadimitriou, who has been selected as a Siebel Scholar, providing him with one year of financial (tuition plus stipend) support. These scholarships are funded from an endowment set up by the Siebel Corporation.

RECENT PUBLICATIONS

… continued from pg. 3

On the Effectiveness of Rate Limiting Mechanisms

Wong, Bielski, Studer & Wang

CMU Parallel Data Lab Technical Report CMU-PDL-05-103, March 2005.

One class of worm defense techniques that received attention of late is to "rate limit" outbound traffic to contain fast spreading worms. Several proposals of rate limiting techniques have appeared in the literature, each with a different take on the impetus behind rate limiting. This paper presents an empirical analysis on different rate limiting schemes using real traffic and attack traces from a sizable network. In the analysis we isolate and investigate the impact of the critical parameters for each scheme and seek to understand how these parameters might be set in realistic network settings. Analysis shows that using DNS-based rate limiting has substantially lower error rates than schemes based on other traffic statistics. The empirical analysis additionally brings to light a number of issues with respect to rate limiting in practice. We explore the impact of these issues in the context of general worm containment.

… continued on pg. 6

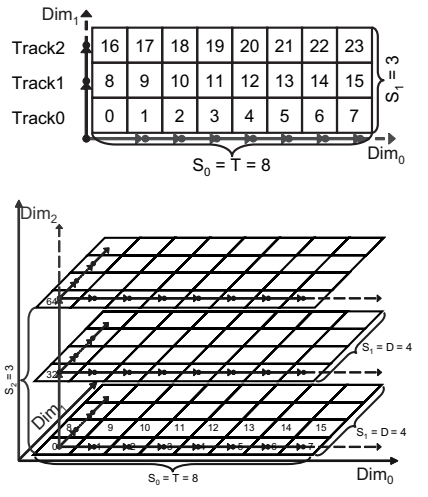
… continued from pg. 5

Multimap: Preserving Disk Locality for Multidimensional Datasets

Shao, Schlosser, Papadomanolakis, Schindler, Ailamaki, Faloutsos & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-05-102, March 2005.

MultiMap is a new approach to mapping multidimensional datasets to the linear address space of storage systems. MultiMap exploits modern disk characteristics to provide full streaming bandwidth for one (primary) dimension and maximally efficient non-sequential access (i.e., minimal seek and no rotational latency) for the other dimensions. This is in contrast to existing approaches, which either severely penalize non-primary dimensions or fail to provide full streaming bandwidth for any dimension. Experimental evaluation of a prototype implementation demonstrates MultiMap's superior performance for range and beam queries. On average, MultiMap reduces overall I/O time by over 50% when compared to traditional naive layouts and by over 30% when compared to a Hilbert curve approach. For scans of the primary dimension, MultiMap and naive both provide almost two orders of magnitude higher throughput than the Hilbert curve approach.



Above, a 2-D dataset is mapped to disks. The first dimension is mapped to the track; the second dimension is mapped to the sequences of the first-α neighbors. Below, a 2-D dataset is mapped to disks. Each surface is a 2-D structure as in the 2-D map. The third dimension is mapped to the sequences of 4th-α neighbors.

Ursa Minor: Versatile Cluster-based Storage

Ganger, Abd-El-Malek, Cranor, Hendricks, Klosterman, Mesnier, Prasad, Salmon, Sambasivan, Sinnamohideen, Strunk, Thereska & Wylie

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-05-104, April 2005.

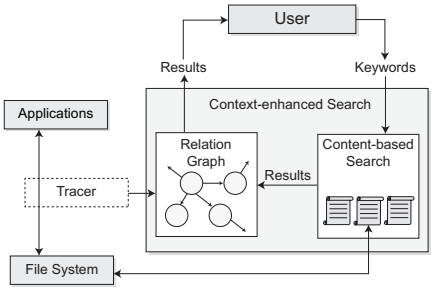
No single data encoding scheme or fault model is right for all data. A versatile storage system allows these to be data-specific, so that they can be matched to access patterns, reliability requirements, and cost goals. Ursa Minor is a cluster-based storage system that allows data-specific selection and on-line changes to encoding schemes and fault models. Thus, different data types can share a scalable storage infrastructure and still enjoy customized choices, rather than suffering from "one size fits all." Experiments with Ursa Minor show performance penalties as high as 2–3X for workloads using poorly-matched choices. Experiments also show that a single cluster supporting multiple workloads is much more efficient when the choices are specialized rather than forced to use a middle-of-the-road value.

Connections: Using Context to Enhance File Search

Soules & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-05-105, April 2005.

The continued growth of personal file systems demands a shift from manual file organization to effective on-demand search tools. Today's best search tools use content analysis techniques to provide targeted, ranked results for user queries. However, these tools are missing a key way that users remember and search for their data: context. Context is the set of external events that a user associates with a file's use: the user's current task, other files being accessed, the time of day, etc. This paper presents Connections, a search system that combines content analysis with context information using temporal locality of file accesses. Through this combination, Connections improves both the false-negative rate (recall) and false-positive rate (precision) over content analysis alone. That is, by adding context information, our system finds more of the desired files and ranks them more accurately.



Architecture of Connections. Both applications and the file system remain unchanged, as the only information required by Connections can be gathered either by a transparent tracing module or directly from existing file system interfaces.

Andy Klosterman presents The Client/Metadata Interface of Ursa Major at last year's Retreat.