Research Statement--An-I Andy Wang

The current management of storage and network data flows resembles the plumbing industry 200 years ago, with limited interchangeable parts and poorly understood interactions between parts. These problems present many research possibilities, and I have recently explored five areas: (1) energy-efficient storage, (2) wireless coordination of mobile computers, (3) memory-based storage, (4) techniques to analyze distributed states, and (5) time-constrained data request handling.

One theme of my research involves manipulating data flow paths and storage to yield benefits. For example, by redirecting requests to data replicated at unused portions of disks, we can create opportunities to power down disks and save energy. By directing data flows according to laws of nature such as electric field lines, we can form disjointed data paths among computers with local knowledge and no explicit coordination. By throttling data flow to a storage component, we can control its timing guarantees. By using specialized data paths to memory and disk storage, we can improve performance significantly. General lessons also help to formulate the vision and requirements for designing and building an overarching storage data path framework.

Another theme concerns combining empirical, simulation, and analytical techniques to probe blind spots in existing system designs and performance characteristics. For example, with my new analytical technique, I can construct models that remove 99.9999% of redundant and unreachable system states. This technique is applicable to validating simulations and characterizing the behaviors of distributed systems with an exponential state space. Also, we were able to use measured timing guarantees to retrofit implementations into theoretical framework. The envelope bound of timing guarantees was actually tighter than the theoretical bound.

To illustrate these themes, I will describe my recent research projects. The first three show how manipulations of data paths can yield benefits in energy savings, performance improvement, and scaling of distributed systems. The remaining two illustrate how the combination of analytical, simulation, and empirical approaches can lead to greater insights into system behaviors. These descriptions are followed by my views on future directions of systems design and performance measurement.

Energy-efficient storage: Power consumption of disks is a major concern for data centers and for places with limited power infrastructure. Designing an energy-saving storage system is difficult because of the peak performance and reliability requirements. We have designed and built one of the first power-saving storage prototypes, PARAID which can achieve a 34% savings in power while meeting performance and reliability constraints. Basically, PARAID exploits unused storage to create opportunities to power down disks and reuses the existing infrastructure to retain reliability and peak performance. Also, PARAID exploits cyclic workload patterns to reduce the number of power cycles to prolong the lifespan of disks. This experimental system identified complex transformations of workloads along the storage data path and unveiled difficulties in evaluating energy-saving systems. It led to two on-going masters' theses. PARAID was published at USENIX FAST, the flagship conference in storage. As one of the top seven papers in the conference, the paper was invited to be published in ACM Transactions on Storage.

Mobile wireless networks: Coordinating mobile computers is challenging because the global view of all computer nodes is often out-of-date and unavailable. One problem concerns constructing disjoint data flows or routes through mobile wireless networks to improve reliability. Inspired by the shapes of electric-field lines, where two poles are connected by disjoint field lines, I invented electric-field-based routing (EFR). With EFR, forming disjoint routes requires no explicit coordination. To forward data, any node needs to know only its neighbors and its position relative to the source and destination. Additionally, each node does not need to track its participation in various routes, since routes are determined on-the-fly based on current neighbors. The lack of coordination and route tracking makes EFR scale well. EFR was published in ACM SIGMOBILE Mobile Computing and Communications Review.

In another piece of work in this area, we improved the popular NS-2 simulation to evaluate mobile wireless networks in metropolitan settings. Notably, we used street maps, experimented with different vehicle mobility models based on traffic rules, validated radio attenuation models due to buildings, and explored the roof-top network infrastructural support. The results were published at IEEE ICCCN and ACM/IEEE MSWiM.

Disk-memory hybrid file system: Disks are cost-effective in capacity; memory is cost-effective in performance. I designed, implemented, and measured Conquest, a file system that combines the strengths of both media. Conquest matches user behaviors and the characteristics of storage. In essence, small files are accessed frequently and are stored in memory. Large files consume the most space and are stored on disk. By creating two specialized data paths to disk and to memory, Conquest can outperform the-state-of-art systems by up to a factor of two. The Conquest design also reflects economic principles of specialization and trade, which can be extended to form storage solutions in the distributed domain. This work was published in IEEE HotOS, the USENIX Annual Technical Conference, and ACM Transactions on Storage.

Analysis of optimistically synchronized data replication systems: Optimistic replication is widely deployed (e.g. airline reservation systems) and allows concurrent updates to data replicas. Although conflicting updates are shown to converge empirically and in simulators, no theoretical understanding of such systems exists beyond two replicas. The challenge is track all update/conflict relationships for each pair of replicas and all transitions from one system state others, depending on how updates propagate. Beyond two replicas, the system states grow exponentially. I invented permuted states, a representation to eliminate 99.9999% of redundant and unreachable system states. For the first time, system designers can enumerate and validate simulation states up ten replicas, and understand how cyclic workload patterns enable such systems to operate stably with low overhead. This work is published at ACM SIGMETRICS, SCS SPECTS, and SCS Simulation Transactions of the Society for Modeling and Simulation International.

Fitting real-time theories into I/O systems: Although some operating systems support applications with timing or real-time constraints, we cannot apply current scheduling theory in a straightforward manner to explain the timing behaviors of I/Os. Basically, the management aspect of data path flow and storage hardware can interfere with timing constraints of I/Os. For example, a disk can handle multiple requests at a time, but it serves them in the order that is convenient or efficient for the disk without

distinguishing requests with real-time constraints. Our solution is to reduce the number of requests handled by such components, whenever deadlines are in jeopardy. In other words, the boundary of control granted to such components can be adjusted, depending on the laxity of timing constraints. Therefore, we can avoid conservative measures if deadlines are usually safe, while meeting deadlines when unsafe situations arise. This disk throttling framework is published at IEEE RTAS. Currently, one Ph.D. student is preparing a prospectus in this area, which will be submitted as a NSF proposal.

For network I/Os, we also developed a theoretical analysis technique based on measured timing characteristics, to derive achievable timing guarantees for workloads that do not fit conventional theoretical models. The result is published at IEEE RTAS.

Future research direction: Although my research areas already span energy savings, distributed coordination, performance improvements, real-time guarantees, and system analysis, I am constantly seeking to expand my interests. With my experience in various aspects of storage, my eventual goal is to design storage data path primitives to build an overarching framework, which can unlock the possibilities of exploring how data flow transformations can be decomposed, combined, and reorganized to meet diverse and even unforeseen goals. The research challenges involve characterizing individual storage components, understanding their interactions, and providing ways for each component to communicate and provide guarantees.

Overall, I have published 1 book chapter, 4 journal papers and 18 conference and workshop papers, at venues such as USENIX, USENIX FAST, ACM TOS, ACM SIGMOBILE, ACM SIGMETRICS, ACM MSWiM, IEEE RTAS, IEEE HotOS, IEEE MASCOTS, SCS Simulation, etc. Of these, nine had an acceptance rate below 25%. In terms of trend, eight papers were published within the last year, reflecting the publication delay due to the need for system implementations. Additionally, one paper is under submission, and two are in preparation. As for non-self citations, I have tracked down 129 unique papers citing my work.

In terms of external funding, I have acquired a $450,000 NSF grant as the lead PI, which funds the energy-efficient storage research. I also acquired a $550,000 NSF grant as a Co-PI, which funds the real-time device driver research. Additionally, I have attracted five students carrying grants from DoE, MCI, Harris, and other scholarships. Further, I have obtained $26,000 internal seed grants to explore research possibilities.

Table 1: Publications, Grants, and Presentations.

	International	National	Regional	State	Total
Refereed Journal Articles	4				4
Invited Book Chapters	1				1
Invited Conference Papers and Proceedings	18				18
Technical Reports				9	9
Other Non-Refereed Publications	4				4
Total Publications	**27**			**9**	**36**
Total Grants		**$997,324**		**$24,999**	**$1,022,323**
Invited Presentations		9		6	15
Refereed Conference Paper Presentations	17				17
Non-Refereed Presentations	2				2
Total Presentations	**19**	**9**		**6**	**34**