Response to the OptIPuter Year 1 Project Review

NSF ITR Cooperative Agreement ANI-0225642

Larry Smarr, Principal Investigator
Cal-(IT)^2, University of California, San Diego
smarr@cs.ucsd.edu

SUBMITTED December 22, 2003

This is the OptIPuter team's response to the NSF Project Review's report following its visit September 29, 2003.

1. **The OptIPuter project's problem space is interesting and has an outstanding team of talented participants. (a) Reviewers voiced concerns that the project/funding is spread thin across too many organizations and researchers who are focused in different directions. (b) By reducing scope and reallocating resources, or minimally making sure that a few *key* research problems are addressed, the project will greatly enhance its opportunity for success and impact. Clearly identify key objectives.**

(a) It should be noted that OptIPuter funding is not consistent throughout the five years of the award, but was below the norm in year 1 and will be above the norm in year 2, pushing some of our deliverables to future years. (This was made clear in our revised Statement of Work, submitted with a revised budget, in Summer 2002.) While we are working on a wide set of OptIPuter components – middleware, protocols, visualization, optical signaling, security, performance, and the applications that drive them – we are planning on combining and integrating these component efforts into a comprehensive software system architecture for the OptIPuter.

(b) We intend to keep our layered management structure intact; however, we agree with the reviewers that defining 3-4 ***research topics*** each year, in order to provide quantitative metrics for evaluating the success and impact of the project, is a good approach. We are now working with the team members to define these research focus areas for year 2, proposed below:

- How to we control lambdas and how do protocols influence their utility?
- How is a LambdaGrid different from a Grid in terms of middleware?
- How can lambdas enhance collaboration?
- How are applications quantitatively helped by LambdaGrids?

These four areas represent goals to help focus OptIPuter team research efforts in year 2 (and beyond). Hence, our teams must review their deliverables in the Cooperative Agreement for year 2 and in the Annual Progress Report and Program Plan (PPP) (pp. 33-38) and then hone their activities to make sure these topics are addressed. We do not advocate changing the deliverables, but tailoring them; i.e., making them more specific.

This will happen by the OptIPuter All Hands Meeting (AHM) in January 2004. At the AHM, team members get to interact with one another, and teams get to interact with other teams. At that time, the discussions need to focus on making the pieces fit together. We propose the following timeline:

- Distribute research focus topics as soon as NSF approves the Year 2 Program Plan
- Have individuals refine year 2 deliverables (described in the Cooperative Agreement and PPP)
- Have individuals send refined deliverables to Team Leaders by the end of December
- Have Team Leaders and their teams review and summarize at the AHM
- Finalize at AHM and submit to NSF shortly thereafter

2. **(a) Reviewers gave positive appraisal of the visualization work. (b) Reviewers were very optimistic that ISI's XCP implementation will yield interesting results, though how it ties in with the rest of the project was questioned. (c) The OptIPuter, in general, relies a great deal on commodity systems. While the work at StarLight, for example, is important in terms of getting operational systems up in the wide area, there is little to no research pursued here (certainly a great deal of difficult engineering). By the same token, the significant focus on getting various vendor systems (Chiaro, Glimmerglass, etc.) up and running will likely not contribute to solving research problems of interest. Reviewers felt that work in these areas should be refocused onto problems that have a research agenda. That is, while the complexity of configuring optical connections is appreciated, there needs to be a higher level context that addresses**

these problems. This context needs to be one which abstracts engineering problems into research problems that are not specific to the technology being used but address problems of interest to the general research community.

(a) We appreciate the reviewers recognizing of our visualization and collaboration efforts. Visualization with and without lambda provisioning will be compared side-to-side by the applications users, and these results will be demonstrated and reported at next year's site visit.

(b) XCP is indeed quite relevant to the OptIPuter project; as reviewers suggest, our presentations may have focused more on infrastructure than the research problems the infrastructure is being designed to address. The OptIPuter project is funding partner USC ISI to customize XCP, a high-performance transport protocol, to meet the requirements of the application drivers in the earth sciences and bioinformatics running on the OptIPuter's underlying *deterministic and schedulable* LambdaGrid." In year 2, ISI plans to explore XCP-enabled Grid applications, such as FTP and (possibly) porting XCP code to the Chiaro optical router. By contrast, ISI was recently awarded NSF Strategic Technologies for the Internet (STI) funding for XCPD, to harden, test and deploy XCP on the commodity *best-effort* Internet on a diverse set of platforms and routers; XCPD will be written up as an Internet specification and submitted to IETF for consideration as a protocol standard. XCPD deliberately excludes any *real science.*

(c) The infrastructure efforts cited, such as StarLight and Glimmerglass, are being put in place to enable lambda management, control and provisioning research, as presented by OptIPuter partner Joe Mambretti of Northwestern University, citing related work by partner UIC partner Oliver Yu and unfunded University of Amsterdam partner Cees de Laat. OptIPuter lightwave control algorithms are being developed using real infrastructure offering extreme bandwidth availability, and are not theoretical models based on simulations. Infrastructure engineering and deployment efforts were critical foundation activities needed to enable the research in the next few years as laid out in the Cooperative Agreement. While Chiaro and Glimmerglass are current vendors of choice, the OptIPuter project is not locked into any particular vendors. In the future, more presentation time will be devoted to the research problems enabled by the infrastructure, rather than the infrastructure itself.

It is important to remember that the OptIPuter is an applications-centric proposal; our overall goal is to develop a new architecture that enables scientists with terabytes and petabytes of data to interactively visualize, analyze, and correlate their data from multiple storage sites connected to optical networks, with precise control. This will remain the driver of our research agenda.

3. **(a) There seems to be a hole (and therefore a risk) in the networking research focus of the project. Specifically, while a number of references were made to the problem of optical provisioning, reviewers felt that greater effort should be put into determining the IP/optical/Ethernet OptIPuter architecture as well as understanding the IP/optical/Ethernet interaction. (b) As the architectural work matures, we expect a much more rigorous investigation of the impact of latency on various architectural choices. (Creating a latency budget for various applications is strongly encouraged).**

(a) Understanding the IP/optical/Ethernet architecture and interaction is perhaps the most difficult research problem – as stated by the reviewers – Infiniband, cluster and storage interconnects all have to play into this as well as the networks. OptIPuter system architecture and visualization/collaboration team members are indeed exploring this issue, and plan to have results to report in year 2. Research challenges dictate that there will several, non-overlapping solutions: some architectures will be more widely accessible, while others will give better performance. We have made significant progress since the site visit in developing broad internal interest in IBM on partnering with us on these research issues. The interest at Telcordia/SAIC remains high as well.

(b) This is an excellent recommendation, which we will begin to pursue in year 2. We will start by doing preliminary back-of-the-envelope calculations for total end-application performance, based on estimates or specs for various contributing latencies for different OptIPuter architectures. End-application performance latency measurements need to be tied to the applications. The UCSD campus network lets us study latency (on the order of less than 0.1ms), while the StarLight/NetherLight network lets us study long-distance latency (less than 100ms). Once the National Lambda Rail (NLR) connects San Diego to Chicago, we can study latency across the USA (on the order of 10ms.) We state in our PPP that UIC, UCSD and SDSU will deploy data management, visualization and collaboration software systems developed in year 1 to OptIPuter application teams, both for testing and for feedback, so the Data/Visualization/Collaboration and Application teams would be the best ones to do some performance studies. These studies will be done over both the campus/Internet2 backbone, and the dedicated optical

networks so we can compare appropriately. The OptIPuter is a five-year project; more quantitative results will come towards the end of the award.

4. **(a) One of the conclusions of the Gigabit Network project is that more effort should have been spent on enhancing the ability of the end hosts to send or receive such bandwidth. The OptIPuter project should avoid the same mistakes, given the assumption of bandwidth growing faster than Moore's Law. Since the building block is the PC, and since bandwidth to the PC is *not* growing faster than Moore's Law, this seems like a critical systems research problem. (b) For example, the UCSD Sun cluster has GigE in each PC. GigE has been available for several years. Will Sun PCs have 10GigE in the next year or two? (c) To avoid repeating history, it would be good for the project to focus on using a cluster to send and sink scalable bandwidth that could be usefully absorbed in a computation. Batch processing of very large datasets to storage does not sound like much research progress.**

(a) Today, the number of PCs with 1GigE connections is vastly greater than the networking can currently service, so we believe that end-host I/O bandwidth is *not* a bottleneck in the foreseeable future. In fact, modern $1000 CPUs run at about 20% utilization when achieving 990Mbps. 10GigE NICs are also available, though costly. We *already* know how to physically absorb such flows onto modern hardware using user-level messaging, co-processed networks and the like (such as active messages, HPVM, PM, etc., which all lead to VIA that is now Infiniband).

Other research groups have measured the process timings of edge devices; see <http://nowlab.cis.ohio-state.edu/projects/mpi-iba/#Results-IA-32>. On the OptIPuter clusters, the processors are faster; UCSD partner Phil Papadopoulos has measured 7-microseconds latency and 829MBytes (6.6-8Gbps achievable) on 4X Infiniband. When 10GigE NICs are routinely available, we have no doubt that we can run at these rates.

The OptIPuter's research focus is first on how to create a working system using parallel paths, prototyped for the next 1-2 years over affordable multi-1GigE paths with 10GigE backbones. Once this system integration research phase is finished we will then push for performance, migrating to multiple 10GigE hosts and striped 10GigE backbones during the out-years of the project. The OptIPuter team assumes that industry will deal with end-host bandwidth requirements and our rapidly developing partnerships with IBM, Sun, and HP are driven by their shared interest in this issue.

Our goal is to understand how applications can fundamentally change, given very large bandwidth. Endpoints go through order-of-magnitude jumps in interface speed, and endpoint bandwidth is approaching memory bandwidth (faster than memory bandwidth is not useful).

(b) The OptIPuter does not favor any specific vendor. There exist 10GigE PC NICs, and while 10GigE Chiaro router cards are expensive, we are also exploring 3D MEMS optical switches (from Calient and Glimmerglass). The UIC-designed "Mod 2" OptIPuter, to be developed in year 2, will have 10GigE cards. We are also exploring new technologies; Fujitsu has a 12-port 10GigE non-blocking Ethernet packet switch chip that will retail for $2,400 per box for copper interfaces, and will come out with a ~$30,000 box with optical interfaces by 2Q2004. We are also seeing a precipitous price drop in the 10GigE space, which we are monitoring closely. These types of systems could become the basis of a campus-area and wide-area OptIPuter, rather than a line card with 4x1 GigE ports. For instance, we are working closely with Chiaro CTO Steve Wallach on when to expect much cheaper 10 GigE interface cards that can be incorporated into the OptIPuter at UCSD.

(c) The OptIPuter's mission is scientific *interaction and collaboration,* to enable scientists to explore massive amounts of previously uncorrelated data. So we agree with the reviewers that batch processing and shipping large data files from one machine to another are best left to the TeraGrid and other Grid production resources.

5. **It would be interesting to develop a model of cost of bandwidth to explore the tradeoffs between local storage and remote access. If one assumes that bandwidth is free but computation and storage costs money, then the solution space is skewed. Certainly, one cannot make tradeoffs if one assumes everything is free. Hence, it is only reasonable to assume that all have costs and perhaps some guidelines might be useful in evaluating some of the solutions and new architecture.**

This is useful feedback. During year 2 we will attempt to develop more quantitative models of the tradeoffs of bandwidth, storage, and computing. In is worth noting that on a 24x7x365 basis, our 10Gb metropolitan link across Chicago from EVL to StarLight costs us under 4 cents a minute to operate and our 10Gb link between StarLight and Amsterdam costs $2 a minute, not counting the already sunk costs of equipment and our metropolitan dark fiber. Four cents a minute is about what a graduate student Research Assistant costs, by comparison, on a yearly basis. It is

also comparable to what two 64-bit PCs with 10GigE NICs, volume visualization graphics cards and displays cost, although these might have a two-or so year lifespan. Thus, we aren't far off from 1-1-1 cost matching of graduate students, 10Gb metro lambdas and current generation computers; as lambdas allow the cost of bandwidth to approach 10% the cost of a graduate student, we believe that networking will no longer be the gatekeeper to our applications' progress. This is the context of the OptIPuter team effort.

Our focus will remain on parallel lightpaths, prototyped over the next year at affordable GigE paths within 10GigE backbones, migrating to multiple 10GigE hosts and striped 10GigE backbones in out years. The process timings of edge devices are an important part of overall performance assessment. We agree that it is important to develop a series of systematic metrics, measurement and instrumentation processes, and analysis tools for the OptIPuter to be able to quantitatively compare its benefits vs. the alternatives. Our Quanta software, for example, has measurement built into every stage; these data is available upon demand to the application developer/user.

6. **Partnerships with domain scientists are strong and positive and will help focus long-term OptIPuter design requirements. In the short term, however, it is not entirely clear that they are demanding much more from the OptIPuter. The value-added to e-science applications could be made stronger. It would be excellent if the OptIPuter could enable something *beyond* the basic demonstration of batch transfer and processing of large datasets. We strongly encourage the OptIPuter team to be creative and explore how new OptIPuter technologies could revolutionize specific e-science applications. If successful, this would be an extremely exciting result for the project.**

During year 1 we succeeded in developing infrastructure and purchasing "Mod 0" equipment for domain scientists in Southern California while Chicago team members developed new tools and technologies, including interactive "Vol-a-Tile" volume visualization software on "Perspec-Tile" tiled displays, and designed "Mod 1" equipment. Much of the OptIPuter environment became operational only a few days prior to the Site Visit. While we agree with reviewers that it is not interesting to ship and process large datasets in batch mode, we need to address this problem in the general grid sense, and then add special abilities for pre-fetching data and rendering terabyte and petabyte data files in close-to-real-time to allow scientists to interactively explore their data and collaborate synchronously.

7. **Given the large number of faculty involved, reviewers were concerned that there were no papers in more prestigious peer-reviewed conferences from ACM, IEEE or OSA. This may be due to the fact that the project is only a year old. Reviewers expressed strong desire to see peer-reviewed journal papers and conference submissions of key output in the future.**

As the reviewers recognize, this project is only one year old. We clearly intend to publish papers and demonstrate our work at major professional conferences; in fact, we have several demos slated for SC 2003. We do have some concern, however, that interdisciplinary research is not typically well received by journals focusing on theory and/or simulations. OptIPuter researchers will certainly aim to publish in relevant journals and major professional conferences.

8. **Because many of the OptIPuter researchers are heavily leveraging other funding sources for their portion of the project, there was some confusion about which efforts are actually receiving substantial project funding. To manage reviewer expectations, we recommend that the next Project Review explicitly include a brief budget overview to explain the funding allocation among participants. (Also, at the next review, please provide hard copies of the presentations at the beginning.)**

Good suggestion. These concerns will be addressed at future NSF Project Reviews.