ADAPTIVE SYSTEMS – PROGRESS REPORT
JULY 2003

RESEARCH PROGRESS

In recent years, explosive growth of the Internet has been a major driving force in the deployment of a variety of wireless technologies. Examples include GPRS, EDGE, CDMA2000, HDR, UMTS. While wireless technologies traditionally were designed to offer circuit-switched voice services, efforts are now underway to integrate mobile radio networks with the Internet to support a rich variety of data applications. Thus far, the focus has been on using only a single wireless interface at a time, and several research challenges related to such use have been explored. Given the scarcity of bandwidth in wireless domain, it's often the case that no single interface can support data intensive applications like video telephony/conference, telemedicine, and large file transfers. Even though the 3G radio access technology UMTS is touted to provide up to 2Mbps, this is under the best of conditions, where the channel is dedicated to a single user and the user has excellent radio conditions. In practice, the user is likely to get only a few hundred kbps. This is inadequate - even a low quality 64kbps rate controlled Variable Bit Rate (VBR) interactive video, due to a huge difference between peak to mean rates (about 15), needs at least 300kbps capacity to achieve adequate quality reception.

Restricting wireless technology use to support only a subset of applications that require low bandwidths less appealing and difficult for operators to gain returns on the huge investments made in acquiring license to spectrum. However, when coverage areas of the different wireless technologies overlap, there is no need to restrict oneself to a single interface. The simultaneous use of multiple interfaces permits bandwidth aggregation, thereby allowing support for demanding applications that need high bandwidths. Further, this can have additional advantages for some applications in terms of increasing reliability, where some or all packets can be duplicated and sent on the multiple interfaces. Also, it can help in mobility management, where the delay associated with handoff can be significantly reduced when an alternate communication path exists.

There are several issues that need to be addressed to enable simultaneous use of multiple interfaces. These span different layers of the protocol stack and need multi-disciplinary approaches to devise appropriate architecture and scheduling policies. We have looked in depth at the problem of bandwidth aggregation. Since TCP is the dominant transport protocol in use today, we have evaluated the performance of a file transfer using TCP over such bandwidth aggregation. We have also looked at the performance of real-time interactive applications with strict QoS requirements.

Towards realizing the objective of simultaneous use of multiple interfaces, we have proposed a general framework in the form of an architecture based on an extension to Mobile-IP, a popular network layer mobility management protocol standardized by IETF. In contrast to transport/application layer solution, such a network layer approach introduces minimal changes to the infrastructure while handling mobility well.

While the use of multiple interfaces allows us to increase throughput, the varying characteristics of the different paths (corresponding to different interfaces) introduce problems in the form of packet reordering. With respect to TCP, this reordering can degrade application throughput significantly as TCP misinterprets reordering as indicative of packet loss and invokes congestion control cutting down the sending rate. This could even counter any gains that can be had through bandwidth aggregation. Therefore, to improve overall performance of TCP, we took a two-pronged approach: (1) we proposed a scheduling algorithm that partitions traffic onto the different paths (corresponding to each interface) such that reordering is minimized; (2) a buffer management policy was introduced at the client to hide any residual reordering from TCP. Simulation studies have shown that this approach can achieve good bandwidth aggregation

under a variety of network conditions. The performance is comparable to "MTCP", an application layer solution that opens multiple TCP connections with one on each interface (the best possible).

For real-time interactive applications, packet reordering introduces excess delay. Because of stringent QoS requirements, this late arrival of packets is often equivalent to loss. Thus, to reduce the delay associated with reordering, we proposed a scheduling policy – Earliest Delivery First (EDF) that partitions the traffic onto different paths such that the QoS requirements of the application are met. We theoretically analyzed the performance of EDF and show that it performs close to an idealized Single Link (SL) discipline, where the multiple interfaces are replaced by a single interface with the aggregate bandwidth. We carried out simulations using video and delay traces and we observed that under a variety of network conditions, EDF mimics SL closely and outperforms by a large margin other straight-forward scheduling policies like weighted round-robin.

While our approach has been based on simulation and analysis so far, we plan to implement some of the ideas on an actual testbed. In addition, we will look into other aspects of bandwidth aggregation related to power consumption, security, etc. While use of multiple interfaces may be a drain on the battery, the power savings in time had through bandwidth aggregation need to be quantified. Security is another important issue in wireless networks. Use of multiple interfaces means securing a greater number of paths, but through proper encryption, connection can be made inherently more secure as it is difficult for an attacker to snoop on all the paths.

We have recently characterized workloads in Public-Area Wireless Networks (PAWNs), and have shown that: (1) user loads are often time varying and location-dependent; (2) user load is often unevenly distributed across access points (APs); and (3) the load on the APs at any given time is not well correlated with the number of users associated with those APs. Administrators in such networks thus have to address the challenge of unbalanced network utilization resulting from unbalanced user load, and also guarantee its users a minimum level of quality of service (e.g., sufficient wireless bandwidth).

We have addressed the challenges of improving PAWN utilization and user bandwidth allocation using a common solution: dynamic, location-aware adaptation. By adaptively varying the bandwidth allocated to users in the wireless hop within certain bounds, coupled with admission control at each AP, the network can accommodate more users as its capacity changes with time. Further, by adaptively selecting the AP that users associate with, the network can relieve sporadic user congestion at popular locations and increase the likelihood of admitting users at pre-negotiated service levels.

We present the problem of first-hop wireless bandwidth allocation as a special case of the well-known online load balancing problem, and have proved that the general online problem of finding an optimal assignment of users to APs in an arbitrary network with arbitrarily sized user bandwidth requests is NP-complete. We therefore developed three online heuristic algorithms for first-hop bandwidth allocation. We describe how these algorithms enable the network to transparently adapt to user demands and balance load across its access points.

Finally, we have evaluated the effectiveness of these algorithms on improving user service rates and network utilization via simulation, incorporating real workloads from campus, conference, and corporate environments. We show that our algorithms improve the degree of balance in the system by over 45% and allocate over 30% more bandwidth to users in comparison to existing schemes that offer little or no load balancing.

Standard video coders often use the immediate past frame as a reference frame with motion compensation for video encoding. In our research, we have used dual reference frame motion compensation in the context of high bandwidth to low bandwidth switching such as from an Ethernet connection to a GPRS system. The implementation is based on MPEG-4. Simulation results show that there is a significant gain in the PSNR for relatively static video sequences.

To evaluate the effectiveness of the dual frame buffer technique, we simulated it by modifying the standard MPEG-4 coder. We allocated additional memory for the long-term frame. An extra bit is transmitted per inter coded MB to inform the decoder which frame it referenced. The intra refresh period was set to 100. Lowering the intra refresh period enhanced the performance of the dual frame encoder, but frequent intra refresh results in higher bit rates, which would exceed the bit rates available for a GPRS system. As inputs, we used the News, Container and the Foreman sequences. To investigate the effects of switching to different low bandwidth networks, we simulated switching from 1 Mbps to low bandwidth networks ranging from 10 kbps (GPRS) to 150 kbps (1xRTT CDMA). Each sequence was encoded employing our dual frame buffer coder as well as by a conventional MPEG-4 coder for comparison.

We found that retaining the high quality frame to be used as the long term past frame for the dual frame encoder results in better video quality as quantified by the PSNR of the decoded sequence at a small cost in memory to retain the dual reference frame.

In this sub-project, we explored end-to-end loss differentiation algorithms (LDAs) for use with congestion-sensitive video transport protocols for networks with either backbone or last-hop wireless links. As our basic video transport protocol, we used UDP in conjunction with a congestion control mechanism extended with an LDA. For congestion control, we used the TCP-Friendly Rate Control (TFRC) algorithm. We extended TFRC to use an LDA when a connection uses at least one wireless link in the path between the sender and receiver. We then evaluated various LDAs under different wireless network topologies, competing traffic, and fairness scenarios to determine their effectiveness. In addition to evaluating LDAs derived from previous work, we also proposed and evaluated a new LDA, ZigZag, and a hybrid LDA, ZBS that selects among base LDAs depending upon observed network conditions.

We evaluated these LDAs via simulation, and found that no single base algorithm performs well across all topologies and competition. However, the hybrid algorithm performed well across topologies and competition, and in some cases exceeded the performance of the best base LDA for a given scenario. All of the LDAs were reasonably fair when competing with TCP, and their fairness among flows using the same LDA depended on the network topology. In general, ZigZag and the hybrid algorithm were the fairest among all LDAs.

We are generally interested in the problem of optimizing the timing and duration of sleep states on mobile devices, with the objective of minimizing power with respect to a QoS constraint. In our current model, there are two power consumption modes, sleep and active. There is associated rate of power consumption at the mobile terminal in each state, as well as a fixed energy consumed in transitioning between the two states. The QoS parameter we are focused on is average delay.

To help gain a better understanding of the general problem, we considered a simple model where there is a single transmitter and receiver. The receiver is the mobile node whose mode we wish to control. The transmitter can give commands to the receiver regarding its sleep state, and forwards incoming streaming data to the receiver appropriately. We formulated this as a Markov decision process, and solved it numerically using dynamic programming. The solutions from the numerical calculations strongly suggest that the optimal policy (that which minimizes average power consumption subject to an average delay constraint) is such that the transmitter should only command the receiver to sleep when there is no data queued at the transmitter. The system thus behaves as a single server queue with vacations. We were able to derive closed form expressions for the optimal sleep duration, as well as the associated minimal rate of power consumption. Future work will be focused on more elaborate models involving multiple users.

In this project, we proposed to study techniques to transform and deliver diverse content adaptively so as to minimize the energy consumed by the wireless handheld. Besides the

content transformation techniques themselves, we had to develop fast and accurate handheld energy modeling techniques that can be used to drive the content transformation algorithms.

In terms of energy modeling, we extended the energy model developed earlier for Hyper-Text Transport Protocol (HTTP)-based textual data communication [i], to include *multimedia access*, specifically image access, and to consider *dynamic channel variations*. In addition to the energy spent in communicating the data, processing multimedia data at the handheld can constitute a significant portion of the total energy consumption. Hence, the computational energy model is extended to include the energy consumed in decompressing images. Next, we considered the effect of other image compression specific parameters as the inputs to the energy model in addition to the size of the objects accessed, as specified in our earlier model. Additionally, channel condition variations can affect the energy consumption significantly. For instance, when the channel condition, represented by the Signal-to-Noise Ratio (SNR) degrades, the energy consumed in accessing the multimedia object increases. This is due the fact that the transmission power used and the number of retransmissions required to communicate increases under poor channel conditions.

In order to consider the above enhancements, we measured energy consumption under different controlled conditions using our data acquisition platform, presented in [i]. Based on the measured data, we performed regression analysis to develop the energy model considering different input variables identified before. The input parameters that we considered are the service specific parameters (volume of data requested, image size, image compression parameter) and network-related parameters (SNR). The resulting energy model is validated and used with adaptive image shaping techniques developed in a separate project. The use of the energy model in guiding the adaptive image shaping technique leads to significant savings in energy consumption with minimal degradation to image quality. For example, for a medical image of size 160x160, our adaptive image delivery techniques consume 6.4J of energy compared to 11.6J without adaptation in Palm.Net access network, while degrading image quality from 40.3dB to 32.1dB. After studying the effect of network conditions on energy consumption, we have started to investigate the effect of another network condition, i.e., error patterns on application data. Below, we briefly describe the goal of the current efforts and the architecture proposed.

One significant bottleneck in enabling high quality multimedia applications is the dynamic error condition caused by wireless channel variation. In order to address the erroneous channel conditions, several link and physical layer error control techniques have been proposed to counter the presence of errors by introducing redundancies in transmission. However, the low layer techniques have high associated communication costs in terms of energy and latency due to the added data transmission. Additionally, these physical/link level techniques are oblivious to application requirements such as inherent error resiliency in application data.

We believe that the overhead of error control can be reduced significantly by adapting the error control mechanisms based on current context represented by *application properties* and *wireless channel conditions*. We are developing a *Context-aware Error Control* (*ConECt*) framework that uses application level information to enable low-cost error control through proper selection and configuration of error control mechanisms. *ConECt* would enable trade-off between an application's Quality of Content (QoC) and communication cost in diverse wireless contexts.

Our approach to achieve the above goal consists of two functional steps: (1) characterization of the effect of different error profiles and error control mechanisms on application's data in terms of Quality of Content (QoC) and communication cost (energy/latency/access cost), and (2) dynamic selection of the error control mechanism at runtime using the pre-characterized application's error-effect models developed in the first step. The proposed *ConECt* framework uses current channel conditions to select appropriate error control mechanism in order to reduce communication cost without affecting QoC significantly. The framework is compatible with current wireless standards as it chooses an error control mechanism from the set of choices allowable by the standard specifications. For example, the majority of 3G wireless data standards currently have support for multiple channel coding algorithms (i.e. convolutional or turbo), and different coding rates at the physical layer [ii]. As the *ConECt* framework does not require any modification

to the existing functionality, it can be deployed easily on current wireless data networks. We are currently evaluating the above framework under diverse channel conditions and data types.

---

i D.Panigrahi, A.Raghunathan, G.Lakshminarayana, S.Dey, "Energy Modeling for Wireless Internet Access", in Proc. Intl. Conf. on Third Generation Wireless and Beyond, pp.332-337, San Francisco, May 2001

ii Eduardo Esteves, Peter J. Black and Mehmet I. Gurelli, "Link Adaptation Techniques for High-Speed Packet Data in Third Generation Cellular Systems", in Proceedings of European Wireless Conference, 2002