Definition of a Mobile Context

NB. A mobile device context can be characterised by the following properties:

a. The number of distinct packets transmitted must be kept minimal, but packet size is less important – it is far cheaper to send one 300kb packet than 10 30kb packets in succession.
b. Optimal power usage in a mobile device is important for transmission but not so much for reception. This asymmetry can greatly affect many optimisations.
c. The device is not always connected, so an always-on perspective is not feasible at a detailed (‘immediate’ response or ‘real-time’) time level. On the other hand, a mobile user is almost always reachable in a less detailed (delayed response) time level. This makes cache invalidation reports a feasible alternative to propagation where updates are made without first checking if necessary.
d. In a mobile-mesh scenario, where all devices share one or a few base stations, there is significant cost involved if a device changes base-station to one not in the existing set.
e. Some mobile communication is proximity-based, whether by sharing base stations or the devices are physically close to each other. This could be most simply modelled as two nodes in a network suddenly communicating - prompted only by influences outside of the network (eg a random distribution of communication paths apparently independent of network topology). Note that there are better proximity models than this – some of which utilize the network topology.
f. Server broadcasting is *much* cheaper than one-to-one client-server communication if many devices need to be updated with common data. But limited bandwidth can severely limit the amount of data that should be broadcast. (How is full-duplex communication handled?)
g. Typical mobile browsing behaviour is different to typical desktop browsing behaviour. Mobile browsing tends to be more application-targeted and involve fewer different web sites than typical desktop usage, but involve this subset of web sites more frequently. This subset often contains material targeted specifically to mobile devices. However, content is more likely to be dynamic rather than static (caching implications are…?).
h. Micro-payments are feasible, secure and affordable in a mobile-phone context, but not so much in a desktop environment. This can significantly change website topology and design (think of downloadable ring-tones – they are rarely a single click away).

Guide to Questions

Unlike previous occasions, where complete references were provided to cover each question, here many of the questions will need to be researched more openly. The questions themselves are more speculative and require you to consider your knowledge and experience as a web user. We emphasize more strongly the degree of information filtering that you need to display in obtaining the relevant information (and properly citing these). Typically, 3-4 main references are expected to support or refute your argument.

To assist you, with the approach, we will go through the following question:

Describe Belady’s Anomaly. How would this affect data cache usage in a mobile device context?
In point form:

- Define Belady’s Anomaly from general references
  - Look for review papers on the general area. You may not see Belady in papers in the mobile area (why?)
  - Determine when the anomaly applies and to what forms of activities
    - Anything other than mobile? Than caching?
  - Ask “to what degree is this) or can it be made) relevant to mobile caching?”
  - Are the assumptions made valid in our case?

- Find the original paper(s) describing this problem
  - Follow the citations. Many papers citing the Belady paper will also cite others on related problems.
  - See how the assumptions above are addressed by these other papers (for example, the FIFO assumption seems pretty restrictive)
  - List the assumptions and restrictions in each case (perhaps in a table). Some may apply to our scenario, some may not, and some may apply under more restricted conditions. For example, the always-on condition may not apply in the mobile context, but we might be able to assume that it does for certain activities.

- Given our table of references, methods and assumptions, we can create a Venn diagram of possible candidates for discussion. Any broken or restrictive assumptions should be separately listed. It is likely that there may be papers addressing these restrictions. Look for them.

- Now invert the list.
  Instead of listing the papers and all their methods and assumptions, we list each assumption in point form and all papers that relate to that assumption.

- We can now write a paper of the form:
  - (introduction)
    In queuing theory, there is an interesting phenomenon called Belady’s anomaly….
    - To describe the effect, let us create an example…
  - (precise definition of problem domain in as abstract form as possible)
    Belady’s Anomaly can be seen in quite a few different situations, for example …
    - 1
    - …mobile caching context…
  - (description of current context in which the problem is to be examined)
    To see how Belady’s anomaly might affect mobile caching, we need to understand precisely when the Anomaly it applies, and how the mobile situation differs from the context in which it was originally described…..
    - .list of assumptions for the anomaly…
    - …list of assumptions for a mobile caching context…
  - (experimental results)
    … not for this assignment
  - (conclusion)
    …so we can avoid the anomalous situation if we ensure that in all cases the following conditions apply:
    - 1
    - …
• Have a friend, not familiar to the domain proof-read. Note his/her questions seeking clarification. Ideally, except for ‘obvious’ material (for which a general reference book can be cited), the reader should not need to go to the cited works to have an overall understanding of the paper.

Questions
Note that these questions vary in difficulty.

1. Belady’s anomaly is best demonstrated with equal sized cache components using a FIFO cache replacement algorithm. On the web this is not realistic. How would the typical size distribution of components on the web affect the algorithm (assume FIFO)?

2. When a server updates data that is already cached within clients, the server can send an invalidation report to inform clients of the update. This can be done either as a broadcast or only to those clients that the server knows are affected. Discuss the advantages of each approach in a mobile context.

3. When a client receives an invalidation report, it may choose to update its cache immediately, or indicate that an update is required prior to subsequent usage, and update only when required. Discuss the advantages and disadvantages of each approach in a mobile context.

4. It would seem obvious to use write-through caches where a client can update cached data. This essentially means that a client sends the equivalent of an invalidation report and the new data to the server. In a mobile context this may not always be ideal – particularly when disconnected. Discuss circumstances where sending both invalidation report and changed data is not ideal, and the risks involved. Can some compromise be reached? Eg. Consider location data.

5. The rise of AJAX usage changes the fundamental web model where there is a 1:1 correspondence between URL and resource, to one where a URL can link to many resources (eg gmail uses separate AJAX transfers for each mail item).

   Caching is typically done at fragment level (fragment caching), not document level, but documents are more independent from each other compared to AJAX fragments to a given document. So what are the implications of AJAX fragment caching in a mobile context?

6. While AJAX subdivides a document into distinct parts that can be cached as fragments, or even as normal files, it is also possible for an intelligent cache to split a single page into static and dynamic components. Describe how such a cache might work. One approach is to use ESI (Edge Side Includes – what is the ‘edge’ referred to?) as proposed by W3, Akamal Inc. and others. This implies that the cache needs hints in order to determine where to split the page. Have there been attempts to automate this splitting without resorting to another language (like ESI)?

7. Distributed caching, in which multiple different servers can cache different parts of a web page, can be implemented using standards such as ICAP (RFC 3507 – “To keep up with the growth in the number of clients, there has been a move towards architectures that scale better through the use of replication, distribution, and caching.”) or ESI as proposed by W3, Akamal Inc. and others. This provides a unique business model where Akamal excel. Describe the business model, and its
implications on mobile caching. What if a mobile carrier (such as Optus) were to provide such caching? Are there security/privacy concerns here? Compare this approach to fragment caching. Consider the age of the standards.

8. As http://www.thinkvitamin.com/features/ajax/5-ways-to-optimize-ajax-in-ruby-on-rails shows for Ruby on Rails and http://cs.rthand.com/blogs/blog_with_righthand/pages/Using-page-fragment-caching-and-AJAX-thing.aspx for ASP, AJAX caching can be done at various places: server (for generated content); external proxy cache; browser cache; and manual JavaScript cache (eg mouse-over images). Under what circumstances is each of these cache locations likely to produce the best improvement when optimised? What might be the best global scenario? The worst? These optima may be data-dependent. Discuss some generic data display classes (eg newspaper, blog, picture gallery style) which might yield specific optimisations.

9. ASP.NET offers a number of OutputCache parameter settings to deal with the caching of generated code (see http://asp.net-tutorials.com/caching/introduction/). Discuss the implications of each of these settings in a mobile context. Do not assume that the mobile device knows .NET, but include discussion of any advantages that a .NET awareness might bring. How does this differ from an equivalent Java JSP/Java ME platform?

10. Since a mobile device is not always connected to the net, in the times when it is known in advance that it will be offline for some time, a cache may be pre-emptively filled with the data items expected/predicted to be needed in the future. This process is called data hoarding. Is data hoarding unique to mobile usage, or is it used in other contexts, and how? Give evidence.

11. Mobile peer to peer caching could be improved in a situation where there are crowds of people in close proximity to each other. In this case P2P could be implemented using Bluetooth, WiFi or Broadband. What kind of issues can arise in this situation? How could caching help here? Assume that the people in the crowd are NOT static (topology?) – ie they move!.

12. Mobile data-collection is a typically asymmetric operation that is least efficient in a mobile context if performed without caching (broadcast might be seen to be most efficient). In turn, maximal caching limits real-time response to data-collected, so some form of middle-ground might seem best.

But when multiple data-sources need to be merged, and action taken based on the result, data caching on each device can become very problematic – particularly in a partially connected environment (eg. imposed on motes by power usage restrictions). Discuss how mechanisms such as distributed and targeted invalidation reports generated from either client data sources or the data collection server can help improve this situation.