Subject: Re: 16 bits wchar

Posted by sergei on Tue, 25 Sep 2007 23:56:04 GMT

View Forum Message <> Reply to Message

As much as I'd like to see RTL in U++, I agree that unicode should, if possible, be fixed. RTL is built upon unicode, so a solid base - unicode strings storage - is essential. Who knows, maybe tomorrow someone will need Linear B.

I was thinking of UTF-32 as a possible main storage format. I wrote a simple benchmark to see what are the speeds with the 3 sizes of character. Here are the results (source attached):

```
Size: 64; Iterations: 10000000; 8: 2281; 16: 2125; 32: 2172; Size: 128; Iterations: 5000000; 8: 1625; 16: 1453; 32: 2391; Size: 256; Iterations: 2500000; 8: 1328; 16: 1515; 32: 1578; Size: 512; Iterations: 1250000; 8: 1375; 16: 1141; 32: 1141; Size: 1024; Iterations: 625000; 8: 1172; 16: 953; 32: 984; Size: 2048; Iterations: 312500; 8: 1094; 16: 875; 32: 906; Size: 4096; Iterations: 156250; 8: 1109; 16: 938; 32: 859; Size: 8192; Iterations: 78125; 8: 1110; 16: 890; 32: 922; Size: 16384; Iterations: 39062; 8: 1000; 16: 813; 32: 4047; Size: 32768; Iterations: 19531; 8: 1000; 16: 2250; 32: 3906; Size: 65536; Iterations: 9765; 8: 1656; 16: 2172; 32: 3812; Size: 131072; Iterations: 4882; 8: 1625; 16: 2125; 32: 3782; Size: 262144; Iterations: 2441; 8: 1593; 16: 2110; 32: 3781; Size: 524288; Iterations: 1220; 8: 1563; 16: 2109; 32: 3984;
```

IMHO, 32-bit values aren't much worse than 16-bit. For search/replace operations - non-32-bit values would have significant overhead for characters outside main plane.

Converting UTF-32 to other formats shouldn't be a problem. But what I like most is that character would be the same as cell (unlike UTF-16 which might have 20 cells to store 19 characters).

Edit: I didn't mention that I tested basic read/write performance. UTF handling would add overhead to 8 and 16 formats, but not to 32 format. I also remembered the UTF8-EE issue. UTF-32 could solve it easily. IIRC only 21 bits are needed for full unicode, so there's plenty of space to escape to (without overtaking private space).

## File Attachments

1) UniCode.cpp, downloaded 542 times