Computer Arithmetic 2016–2017 | **Homework 5 Solutions**

1 Patent search

As you discovered, the language used in patents is quite different from that used in papers. The intent of this problem was to let you make this discovery and to teach you how to look for patents since in many cases companies only publish their work in the form of patents. As researchers, you need to know how to find such information and understand it. So, this problem provides you the ability to

1. search patents and research papers,
2. understand their contents,
3. compare them,
4. comment on your findings,
5. implement parts of what you read,
   and
6. evaluate that implementation.

All these actions are necessary for your research career. I hope you benefited.

2 Operations in binary32

In this exercise, care must be given to subnormal numbers.

1. Intermediate result $(A \times B)$ is below **min** but it is carried as a subnormal number so that:

$$(A \times B) \times C = (1).0100 \times 2^{-124}$$

2. No intermediate results below **min**:

$$A \times (B \times C) = (1).0100 \times 2^{-124}$$

   Notice that if we did not have subnormal numbers and a flush to zero was used then the result here remains the same but that of $(A \times B) \times C$ would be zero.

3. Even though the value of A is shifted out of the mantissa and guard bits, it is still retained in the sticky bit and used to round up the last bit of the final result.

$$A + B + C = (1).\underbrace{0000000}_{7}1\underbrace{00000000000000}_{14}1 \times 2^{5}$$

4. Result is still below **max**.

$$C \times D = (1).0100 \times 2^{127}$$

5. Result is above **max** and due to the rounding it is set to the maximum value.

$$(2 \times C) \times D = (1).\underbrace{11111 \cdots 11111}_{23} \times 2^{127}$$

3 Interval conversions

1. When ℓ and u are greater than half the maximum representable number then m will be between them within the representable range. However, $m = 0.5(\ell + u)$ suffers from premature overflow since the sum is larger than the maximum and m is not correctly calculated. The proposed form $\ell + 0.5 * (u - \ell)$ is safer in those cases.

2. A wide interval with ℓ negative, u positive, and their absolute values greater than half the maximum representable number causes an overflow in the calculation of $u - \ell$. Hence, the form $m = 0.5\ell + 0.5u$ and $r = 0.5u - 0.5\ell$ is even better. In fact, as long as both ℓ and u are representable, this last form does not suffer from overflow but it may suffer from underflow. It may underflow due to the multiplication by 0.5 if the value of either ℓ or u is the least subnormal.

3. Four operations are enough to get both m and r.

$$\begin{aligned} a &= 0.5u \\ b &= 0.5\ell \\ m &= a + b \\ r &= a - b \end{aligned}$$

4. Yes, semi infinite intervals such as $[3, \infty[$ are easily representable in inf-sup notation but not in mid-rad. Other cases may lead to wider intervals when converted, for example $[0, +min]$ in inf-sup where $+min$ is the miniumum subnormal number becomes $(0, +min)$ in mid-rad which includes negative numbers. On the other hand, $(+max, 6)$ in mid-rad becomes $[max - 6, \infty[$ in inf-sup which is much wider.
